PROCEEDINGS

The 9th
ISLAMIC COUNTRIES
CONFERENCE
ON STATISTICAL SCIENCES
2007 (ICCS-IX)

12-14 December 2007

Statistics in the contemporary world - theories, methods and applications

ORGANISER

MAIN SPONSOR

PNB
Permodalan Nasional Berhad

PROCEEDINGS

Proceedings of The 9th Islamic Countries Conference on Statistical Sciences 2007
ICCS-IX 12-14 Dec 2007

A State Space Model in Small Area Estimation

Kusman Sadik[1] and Khairil Anwar Notodiputro[2]

Department of Statistics
Bogor Agricultural University / IPB
Jl. Raya Dramaga, Bogor, Indonesia 16680

Abstract

There have been two main topics developed by statisticians in a survey, i.e. sampling techniques and estimation methods. The current issues in estimation methods relate to estimation of a particular domain having small size of samples or, in more extreme cases, there are no sample available for direct estimation (Rao, 2003). There is a growing demand for reliable small area statistics in order to asses or to put into policies and programs. Sample survey data provide effective reliable estimators of totals and means for large area and domains. But it is recognized that the usual direct survey estimator performing statistics for a small area, have unacceptably large standard errors, due to the circumtance of small sample size in the area. In fact, sample sizes in small areas are reduced, due to the circumtance that the overall sample size in a survey is usually determined to provide specific accuracy at a macro area level of aggregation, that is national territories, regions and so on. The most commonly used models for this case, usually in small area estimation, are based on generalized linear mixed models (GLMM). It is happened some time that some surveys are carried out periodically so that the estimation could be improved by incorporating both the area and time random effects. In this paper we propose a state space model which accounts for the two random effects and is based on two equation, namely transition equation and measurement equation.

Key words: direct estimation, indirect estimation, small area estimation (SAE), general linear mixed model (GLMM), empirical best linear unbiased prediction (EBLUP), block diagonal covariance, Kalman filter, state space model.

1. Introduction

The problem of small area estimation is how to produce reliable estimates of area (domain) characteristics when the sample sizes within the areas are too small to warrant the use of traditional direct survey estimates. The term of small area usually denote a small geographical area, such as a county, a province, an administrative area or a census division. From a statistical point of view the small area is a small domain, that is a small sub-population constituted by specific demographic and socioeconomic group of people, within a

[1] kusmansadik@yahoo.com
[2] khairiln@bima.ipb.ac.id

larger geographical areas. Sample survey data provide effective reliable estimators of totals and means for large areas and domains. But it is recognized that the usual direct survey estimators performing statistics for a small area, have unacceptably large standard errors, due to the circumstance of small sample size in the area. In fact, sample sizes in small areas are reduced, due to the circumstance that the overall sample size in a survey is usually determined to provide specific accuracy at a macro area level of aggregation, that is national territories, regions ad so on (Datta and Lahiri, 2000).

Demand for reliable small area statistics has steadily increased in recent years which prompted considerable research on efficient small area estimation. Direct small area estimators from survey data fail to borrow strength from related small areas since they are based solely on the sample data associated with the corresponding areas. As a result, they are likely to yield unacceptably large standard errors unless the sample size for the small area is reasonably large(Rao, 2003). Small area efficient statistics provide, in addition of this, excellent statistics for local estimation of population, farms, and other characteristics of interest in post-censual years.

2. Indirect Estimation in Small Area

A domain (area) is regarded as large (or major) if domain-specific sample is large enough to yield direct estimates of adequate precision. A domain is regarded as small if the domain-specific sample is not large enough to support direct estimates of adequate precision. Some other terms used to denote a domain with small sample size include local area, sub-domain, small subgroup, sub-province, and minor domain. In some applications, many domains of interest (such as counties) may have zero sample size.

In making estimates for small area with adequate level of precision, it is often necessary to use indirect estimators that borrow strength by using thus values of the variable of interest, y, from related areas and/or time periods and thus increase the effective sample size. These values are brought into the estimation process through a model (either implicit or explicit) that provides a link to related areas and/or time periods through the use of supplementary information related to y, such as recent census counts and current administrative records (Pfeffermann 2002; Rao 2003).

Methods of indirect estimation are based on explicit small area models that make specific allowance for between area variation. In particular, we introduce mixed models involving random area specific effects that account for between area variation beyond that explained by auxiliary variables included in the model. We assume that $\theta_i = g(\bar{Y}_i)$ for some specified $g(.)$ is related to area specific auxiliary data $\boldsymbol{z}_i = (z_{1i}, \ldots, z_{pi})^T$ through a linear model

$$\theta_i = \boldsymbol{z}_i^T \boldsymbol{\beta} + b_i v_i, \qquad i = 1, \ldots, m$$

where the b_i are known positive constants and $\boldsymbol{\beta}$ is the px1 vector of regression coefficients. Further, the v_i are area specific random effects assumed to be independent and identically distributed (iid) with

$$E_m(v_i) = 0 \text{ and } V_m(v_i) = \sigma_v^2 \ (\geq 0), \text{ or } v_i \sim \text{iid } (0, \sigma_v^2)$$

3. Generalized Linear Mixed Model

Datta and Lahiri (2000), and Rao(2003) considered a general linear mixed model (GLMM) covering the univariate unit level model as special cases:

$$\mathbf{y}^P = \mathbf{X}^P \boldsymbol{\beta} + \mathbf{Z}^P \mathbf{v} + \mathbf{e}^P$$

Random vectors $\mathbf{v}$ and $\mathbf{e}^P$ are independent with $\mathbf{e}^P \sim \mathbf{N}(\mathbf{0}, \sigma^2 \boldsymbol{\psi}^P)$ and $\mathbf{v} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{D}(\boldsymbol{\lambda}))$, where $\boldsymbol{\psi}^P$ is a known positive definite matrix and $\mathbf{D}(\boldsymbol{\lambda})$ is a positive definite matrix which is structurally known except for some parameters $\boldsymbol{\lambda}$ typically involving ratios of variance

components of the form σ_i^2/σ^2. Further, $\mathbf{X}^P$ and $\mathbf{Z}^P$ are known design matrices and $\mathbf{y}^P$ is the N x 1 vector of population y-values. The GLMM form :

$$\mathbf{y}^P = \begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{X}^* \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z} \\ \mathbf{Z}^* \end{bmatrix} \mathbf{v} + \begin{bmatrix} \mathbf{e} \\ \mathbf{e}^* \end{bmatrix}$$

where the asterisk (*) denotes non-sampled units. The vector of small area totals (Y_i) is of the form $\mathbf{Ay} + \mathbf{Cy}^*$ with $\mathbf{A} = \oplus_{i=1}^{m} \mathbf{1}_{n_i}^T$ and $\mathbf{C} = \oplus_{i=1}^{m} \mathbf{1}_{N_i - n_i}^T$ where $\oplus_{i=1}^{m} \mathbf{A}_u = \text{blockdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_m)$.

We are interested in estimating a linear combination, $\mu = \mathbf{1}^T\boldsymbol{\beta} + \mathbf{m}^T\mathbf{v}$, of the regression parameters $\boldsymbol{\beta}$ and the realization of $\mathbf{v}$, for specified vectors, $\mathbf{l}$ and $\mathbf{m}$, of constants. For known δ, the BLUP (*best linear unbiased prediction*) estimator of μ is given by (Rao, 2003)

$$\tilde{\mu}^H = t(\delta, \mathbf{y}) = \mathbf{1}^T \tilde{\boldsymbol{\beta}} + \mathbf{m}^T \tilde{\mathbf{v}} = \mathbf{1}^T \tilde{\boldsymbol{\beta}} + \mathbf{m}^T \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

Model of indirect estimation, $\hat{\theta}_i = \mathbf{z}_i^T\boldsymbol{\beta} + b_i v_i + e_i, \quad i = 1, \ldots, m,$ is a special case of GLMM with block diagonal covariance structure. Making the above substitutions in the general form for the BLUP estimator of μ_i, we get the BLUP estimator of θ_i as:

$$\tilde{\theta}_i^H = \mathbf{z}_i^T \tilde{\boldsymbol{\beta}} + \gamma_i(\hat{\theta}_i - \mathbf{z}_i^T \tilde{\boldsymbol{\beta}}), \text{ where } \gamma_i = \sigma_v^2 b_i^2 / (\psi_i + \sigma_v^2 b_i^2), \text{ and}$$

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_v^2) = \left[\sum_{i=1}^{m} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\psi_i + \sigma_v^2 b_i^2}\right]^{-1} \left[\sum_{i=1}^{m} \frac{\mathbf{z}_i \hat{\theta}_i}{\psi_i + \sigma_v^2 b_i^2}\right]$$

4. State Space Models

Many sample surveys are repeated in time with partial replacement of the sample elements. For such repeated surveys considerable gain in efficiency can be achieved by borrowing strength across both small areas and time. Their model consist of a sampling error model

$$\hat{\theta}_{it} = \theta_{it} + e_{it}, \quad t = 1, \ldots, T; \; i = 1, \ldots, m$$

$$\theta_{it} = \mathbf{z}_{it}^T \boldsymbol{\beta}_{it}$$

where the coefficients $\boldsymbol{\beta}_{it} = (\beta_{it0}, \beta_{it1}, \ldots, \beta_{itp})^T$ are allowed to vary cross-sectionally and over time, and the sampling errors e_{it} for each area i are assumed to be serially uncorrelated with mean 0 and variance ψ_{it}. The variation of $\boldsymbol{\beta}_{it}$ over time is specified by the following model:

$$\begin{bmatrix} \beta_{itj} \\ \beta_{ij} \end{bmatrix} = \mathbf{T}_j \begin{bmatrix} \beta_{i,t-1,j} \\ \beta_{ij} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} v_{itj}, \quad j = 0,1,\ldots, p$$

It is a special case of the general state-space model which may be expressed in the form

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t; & E(\boldsymbol{\varepsilon}_t) = 0, \quad E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^T) = \boldsymbol{\Sigma}t \\ \boldsymbol{\alpha}_t &= \mathbf{H}_t\boldsymbol{\alpha}_{t-1} + \mathbf{A}\boldsymbol{\eta}_t; & E(\boldsymbol{\eta}_t) = 0, \quad E(\boldsymbol{\eta}_t\boldsymbol{\eta}_t^T) = \boldsymbol{\Gamma} \end{aligned}$$

where $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are uncorrelated contemporaneously and over time. The first equation is known as the *measurement equation*, and the the second equation is known as the *transition equation*. This model is a special case of the general linear mixed model but the state-space form permits updating of the estimates over time, using the Kalman filter equations, and smoothing past estimates as new data becomes available, using an appropriate smoothing algoritm.

The vector $\boldsymbol{\alpha}_t$ is known as the *state vector*. Let $\tilde{\boldsymbol{\alpha}}_{t-1}$ be the BLUP estimator of $\boldsymbol{\alpha}_{t-1}$ based on all observed up to time (t-1), so that $\tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathbf{H}\tilde{\boldsymbol{\alpha}}_{t-1}$ is the BLUP of $\boldsymbol{\alpha}_t$ at time (t-1). Further, $\mathbf{P}_{t|t-1} = \mathbf{H}\mathbf{P}_{t-1}\mathbf{H}^T + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$ is the covariance matrix of the prediction errors $\tilde{\boldsymbol{\alpha}}_{t|t-1} - \boldsymbol{\alpha}_t$, where

$$\mathbf{P}_{t-1} = E(\tilde{\boldsymbol{\alpha}}_{t-1} - \boldsymbol{\alpha}_{t-1})(\tilde{\boldsymbol{\alpha}}_{t-1} - \boldsymbol{\alpha}_{t-1})^T$$

is the covariance matrix of the prediction errors at time (t-1). At time t, the predictor of $\boldsymbol{\alpha}_t$ and its covariance matrix are updated using the new data ($\mathbf{y}_t$, $\mathbf{Z}_t$). We have

$$\mathbf{y}_t - \mathbf{Z}_t \tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathbf{Z}_t(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|t-1}) + \boldsymbol{\varepsilon}_t$$

which has the linear mixed model form with $\mathbf{y} = \mathbf{y}_t - \mathbf{Z}_t \tilde{\boldsymbol{\alpha}}_{t|t-1}$, $\mathbf{Z} = \mathbf{Z}_t$, $\mathbf{v} = \boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|t-1}$, $\mathbf{G} = \mathbf{P}_{t|t-1}$ and $\mathbf{V} = \mathbf{F}_t$, where $\mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t^{\mathrm{T}} + \boldsymbol{\Sigma}_t$. Therefore, the BLUP estimator $\tilde{\mathbf{v}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{y}$ reduces to

$$\tilde{\boldsymbol{\alpha}}_{t-1} = \tilde{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{Z}_t^T \mathbf{F}_t^{-1}(\mathbf{y}_t - \mathbf{Z}_t \tilde{\boldsymbol{\alpha}}_{t|t-1})$$

5. Case Study

Model of small area estimation can be applied to estimate the average of households expenditure per month for each of *m* = 37 counties in East Java, Indonesia. We used Susenas data (National Economic and Social Survey, BPS 2003-2005) to demonstrate the performance of EBLUP resulted from state space models .

Table 1. Design Based and Model Based Estimates of County Means and Estimated Standard Error

County	Design Based (Direct Estimator)		Model Based (Indirect Estimator)			
			EBLUP		$\text{EBLUP}_{(\text{state space})}$	
	$\hat{\mu}_i$	$s(\hat{\mu}_i)$	$\hat{\mu}_i^H$	$s(\hat{\mu}_i^H)$	$\hat{\mu}_i^{ss}$	$s(\hat{\mu}_i^{ss})$
Pacitan	4.89	0.086	3.89	0.062	5.23	0.038
Ponorogo	5.5	0.148	5.83	0.149	5.73	0.132
Trenggalek	5.3	0.135	6.89	0.155	5.65	0.161
Tulungagung	6.78	0.229	7.06	0.215	7.05	0.172
Blitar	5.71	0.132	5.74	0.198	6.12	0.141
Kediri	5.62	0.105	7.09	0.110	6.45	0.091
Malang	5.94	0.128	6.58	0.112	5.19	0.109
Lumajang	5.07	0.119	4.75	0.118	5.74	0.081
Jember	4.65	0.090	4.96	0.126	5.28	0.113
Banyuwangi	5.98	0.142	5.55	0.124	6.15	0.131
Bondowoso	4.53	0.127	4.64	0.092	5.43	0.105
Situbondo	4.67	0.104	5.89	0.085	4.44	0.074
Probolinggo	5.54	0.154	6.07	0.184	7.34	0.186
Pasuruan	6.31	0.151	4.95	0.121	6.39	0.109
Sidoarjo	9.33	0.169	9.46	0.177	8.32	0.123
Mojokerto	6.91	0.160	6.55	0.135	8.25	0.107
Jombang	6.09	0.131	5.06	0.130	5.96	0.091
Nganjuk	5.56	0.125	4.40	0.041	4.87	0.029
Madiun	5.5	0.139	5.16	0.116	5.46	0.121
Magetan	5.52	0.161	4.84	0.145	4.16	0.132
Ngawi	4.89	0.102	4.61	0.097	4.15	0.086
Bojonegoro	5.06	0.093	5.25	0.067	4.50	0.047
Tuban	6.02	0.114	5.75	0.061	6.47	0.046
Lamongan	6.29	0.106	6.47	0.123	5.69	0.065
Gresik	8.49	0.186	9.07	0.167	9.01	0.198
Bangkalan	6.61	0.140	5.69	0.091	7.00	0.076
Sampang	6.32	0.158	7.20	0.150	6.85	0.182
Pamekasan	5.78	0.107	6.10	0.126	5.93	0.109
Sumenep	5.48	0.108	5.76	0.077	5.09	0.032
Kota Kediri	8.01	0.159	7.60	0.157	7.11	0.144
Kota Blitar	7.98	0.191	7.63	0.159	8.51	0.182
Kota Malang	11.14	0.298	12.63	0.273	11.61	0.225
Kota Probolinggo	9.1	0.183	7.68	0.140	10.50	0.153

County	Design Based (Direct Estimator)		Model Based (Indirect Estimator)			
			EBLUP		$EBLUP_{(state\ space)}$	
	$\hat{\mu}_i$	$s(\hat{\mu}_i)$	$\hat{\mu}_i^H$	$s(\hat{\mu}_i^H)$	$\hat{\mu}_i^{ss}$	$s(\hat{\mu}_i^{ss})$
Kota Pasuruan	7.75	0.149	8.09	0.085	8.41	0.072
Kota Mojokerto	9.45	0.204	9.51	0.235	9.01	0.211
Kota Madiun	8.4	0.162	8.33	0.150	7.62	0.196
Kota Surabaya	11.45	0.328	11.81	0.353	11.16	0.321
Mean		0.149		0.138		0.124

Table 1 shows the design based and model based estimates. The design based estimates is direct estimator based on sampling design. EBLUP estimates, $\hat{\mu}_i^H$, used small area model with area effects (data of Susenas 2005) whereas, $EBLUP_{(ss)}$ estimates, $\hat{\mu}_i^{ss}$, used small area model with area and time effects (data of Susenas 2003 to 2005). The estimated standard errors are denoted by $s(\hat{\mu}_i)$, $s(\hat{\mu}_i^H)$, and $s(\hat{\mu}_i^{ss})$. It is clear from Table 1 that the estimated standard errors of mean for the model based is less than the estimated standard error for the estimates design based. The estimated standard error mean of $EBLUP_{(ss)}$ is less than EBLUP.

6. Conclusion

Small area estimation can be used to increase the effective sample size and thus decrease the standard error. For such repeated surveys considerable gain in efficiency can be achieved by borrowing strength across both small area and time. Availability of good auxiliary data and determination of suitable linking models are crucial to the formation of indirect estimators.

7. Reference

- Datta, G.S, and Lahiri, P. (2000). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors (BLUP) in Small Area Estimation Problems, *Statistica Sinica*, **10**, 613-627.
- Pfeffermann, D. (2002). Small Area Estimation – New Developments and Directions. *International Statistical Review*, **70**, 125-143.
- Pfeffermann, D. and Tiller, R. (2006). State Space Modelling with Correlated Measurements with Aplication to Small Area Estimation Under Benchmark Constraints. S3RI Methodology Working Paper M03/11, University of Southampton. Available from: http://www.s3ri.soton.ac.uk/publications.
- Rao, J.N.K. (2003). Small Area Estimation. John Wiley & Sons, Inc. New Jersey.
- Rao, J.N.K., dan Yu, M. (1994). Small Area Estimation by Combining Time Series and Cross-Sectional Data. *Proceedings of the Section on Survey Research Method.* American Statistical Association.
- Swenson, B., dan Wretman., J.H. (1992). The Weighted Regression Technique for Estimating the Variance of Generalized Regression Estimator. *Biometrika*, **76**, 527-537.
- Thompson, M.E. (1997). Theory of Sample Surveys. London: Chapman and Hall.