Area level model approach to small or large area estimation incorporating auxiliary information

Jae-kwang Kim *
Iowa State University, Ames, U.S.A. - jkim@iastate.edu

Zhonglei Wang
Iowa State University, Ames, U.S.A. - wangzl@iastate.edu

Zhengyuan Zhu
Iowa State University, Ames, U.S.A. - zhuz@iastate.edu

Abstract

Combining information from different source is an important practical problem. Using hierarchial area level models, we establish a framework for combining information from different source to get improved prediction for small or large area estimation. The best prediction is obtained by the conditional expectation of the observable latent variable given all available observation. The model parameters are estimated by two-level EM algorithm. Estimation of the mean squared prediction error is discussed.
Sponsored by National Agricultural Statistical Agency (NASS) of US department of Agriculture, the proposed method was applied to the crop acrage prediction problem combining information from three sources: The first source is the June Area Survey (JAS), which is otained by the probability sampling. The second source is from the Farm Service Agency (FSA) data, which is obtained from a voluntary participation of certain programs. The third source is from the classification of the Cropland Data Layer (CDL).

Keywords: Best prediction; Mean squared error estimation; multi-level models

1. Introduction

In recent years, the demand for reliable small area estimates has increased due to, among other things, their growing use in policy making and allocation of governments funds and in regional planning. Thus, small area estimation has been an area of considerable interest in the recent years. Rao (2003) provides a comprehensive overview of the literatures in small area estimation. Perhaps the most famous use of small area estimation is the SAIPE (Small Area Income and Poverty Estimation) program, which is based on the model first proposed by Fay and Herriot (1979), and the U.S. Department of Education allocates annually over $7 billion of funds to counties based on the SAIPE.

Small area estimation improves the direct estimators using auxiliary variables. The auxiliary variables are often obtained from Census or from other external source. A statistical model is used to link the direct estimator and the auxiliary variable. If the model is constructed using the small areas as the analysis unit, the model is called the area level model. If the model is constructed using the individuals as the analysis unit, the model is called the unit level model. The unit level model approach was first considered by Bettesse, Harter, and Fuller (1988) and was extended by You and Rao (2002) which is based on pseudo EBLUP estimator. The unit level model is not applicable when the auxiliary variables are available only in the small area level.

In the U.S. National Agricultural Statistical Service (NASS), June Area Survey (JAS) is annually conducted to obtain state-level predictions of the crop acres for various commodities. In addition to JAS data, we have two additional auxiliary information in the county (or district) level within each states. One is the Cropland Data Layer (CDL) that is obtained from satellite imagery. The other source is the attribute data (578) from Farm Service Agency (FSA). The CDL data is constructed by applying a machine learning classification technique to the satellite imagery. While there is no sampling error in the CDL data, there is certain level of classification error (or measurement error). The attribute data from FSA is subject to coverage error.

To incorporate these auxiliary information into JAS survey estimates, we use two-level model approach to small area estimation. Two-level model is a useful tool for analyzing data with hierarchical structures. Torabi and Rao (2008) considered a two-level model approach to small area estimation in the context of unit level model approach. Up to the knowledge of authors, there is no work on two-level model small area estimation in the area level model approach.

In this paper, motivated from June Area Survey example at NASS, we propose a novel application of the two-level Fay-Harriott model to small area estimation. The paper is organized as follows. In Section 2, the basic setup, including model and parameter estimation, is presented. In Section 3, best prediction under the two-level model is discussed and its mean squared error estimation is also discussed. In Section 4, some computational details are discussed. Application to JAS crop acreage prediction is presented in Section 5.

2. Basic Setup

To simplify the setup, assume that there are two levels of areas in the population. One is the district level (or county level) and the other is state level. For each state h, There are $m_h (\geq 3)$ districts in each state (h). Let Y_{hi} denote the population total of y for district i in state h. From JAS data, we obtain $\hat{Y}_{hi}$ which is unbiased for Y_{hi}, and a vector of auxiliary variables, $\mathbf{X}_{hi}$, which does not suffer from sampling errors. In addition, we have $\hat{V}_{hi}$ available which estimates the sampling variance of $\hat{Y}_{hi}$. Our goal is to obtain a best predictor of $Y_h = \sum_{i=1}^{m_h} Y_{hi}$ using $\mathbf{X}_{hi}$ and $\hat{Y}_{hi}$.

To incorporate $\mathbf{X}_{hi}$ into the prediction of Y_{hi}, we use the following structural model:

$$Y_{hi} \sim f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h) \tag{1}$$

for some parametric model $f_1(\cdot)$ known up to the state-specific parameter vector $\boldsymbol{\theta}_h$ which satisfies

$$\boldsymbol{\theta}_h \sim f_2(\boldsymbol{\theta}_h; \boldsymbol{\zeta}) \tag{2}$$

for some parametric model $f_2(\cdot)$ known up to parameter vector $\boldsymbol{\zeta}$. Model (1) is the level 1 model (within-state model) and Model (2) is the level 2 model (between-state model). Model (1) and model (2) form a two-level structural model for Y_{hi}. Two-level model is very useful in describing the hierarchical structure of the data and will give a good prediction in the state level population parameters. The between-state model in (2) enables to borrow strength from observations outside the states.

In addition to structural model, we need another model, called sampling error model, to incorporate the direct estimate $\hat{Y}_{hi}$ obtained from survey data. The sampling error model can be written

$$\hat{Y}_{hi} \sim g(\hat{Y}_{hi} \mid Y_{hi}) \tag{3}$$

for some known distribution $g(\cdot)$. In many cases, we can assume that $\hat{Y}_{hi}$ follows from normal distribution with mean Y_{hi} and variance $\hat{V}_{hi}$. If the normality is questionable, then a variance stabilizing transformation can be considered. [Remark: We can also consider other parametric distribution or some nonparametric distribution estimated by bootstrap, but it will be discussed in later report.] We further assume that $\hat{Y}_{hi}$ is conditionally independent of $\mathbf{X}_{hi}$ given Y_{hi}. That is,

$$\hat{Y}_{hi} \perp \mathbf{X}_{hi} \mid Y_{hi} \tag{4}$$

Thus, $\hat{Y}_{hi}$ is a surrogate variable for Y_{hi}.

Under the above model setup, given the parameter values, we can apply Bayes formula to obtain a prediction model for Y_{hi} by

$$Y_{hi} \mid X_{hi}, \hat{Y}_{hi} \sim \frac{f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h) g(\hat{Y}_{hi} \mid Y_{hi})}{\int f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h) g(\hat{Y}_{hi} \mid Y_{hi}) \mathrm{d}Y_{hi}}. \tag{5}$$

If f_1 is a normal distribution with mean $\boldsymbol{\beta}_h' \mathbf{X}_{hi}$ and variance σ_{hi}^2 and g is also a normal distribution with mean Y_{hi} and variance v_{hi}, then (5) reduces to

$$Y_{hi} \mid (X_{hi}, \hat{Y}_{hi}, \boldsymbol{\theta}_h) \sim N\left[c_{hi}\hat{Y}_{hi} + (1 - c_{hi})\boldsymbol{\beta}_h' \mathbf{X}_{hi}, c_{hi} v_{hi}\right] \tag{6}$$

and

$$c_{hi} = \frac{\sigma_{hi}^2}{v_{hi} + \sigma_{hi}^2}.$$

For the major crops in the NASS project, after some explanatory data analysis which will be presented in Section 6, we use the following normal models

$$Y_{hi} = \beta_{h0} + \beta_{h1} X_{hi1} + \beta_{h2} X_{hi2} + e_{hi}, \quad e_{hi} \sim N(0, X_{hi1}^{\alpha} \sigma_h^2), \tag{7}$$

where X_{hi1} is the FSA estimate of Y_{hi} and X_{hi2} is the CDL estimate of Y_{hi}. We use $\alpha = 1$ for major crops. We may use $\alpha = 2$ for minor crops. The level 2 model is

$$\boldsymbol{\beta}_h \sim N(\boldsymbol{\beta}, \Sigma) \tag{8}$$

where $\boldsymbol{\beta}_h = (\beta_{h0}, \beta_{h1}, \beta_{h2})'$. The sampling error model is

$$\hat{Y}_{hi} = Y_{hi} + u_{hi}, \quad u_{hi} \sim N(0, v_{hi}) \tag{9}$$

where v_{hi} is known. The assumption works well for major crops, whose median district-level yield is more than 150,000 acres for a specific state. For other minor crops, the above normality does not hold and more tailer-made models need to be considered (which will be discussed in a separate report.)

3. Parameter estimation

We now discuss parameter estimation under the above model. Since Y_{hi} is a latent variable, we can apply EM algorithm to compute the parameters in each level. In level 1, we treat $\boldsymbol{\theta}_h$ as fixed and the following EM algorithm can be used. In the E-step of the EM algorithm, we use the parametric fractional imputation of Kim (2011) to facilitate the computation.

1. E-step: Find the conditional distribution of Y_{hi} given $(\mathbf{X}_{hi}, \hat{Y}_{hi}, \boldsymbol{\theta}_h)$

$$Y_{hi} \mid (\mathbf{X}_{hi}, \hat{Y}_{hi}, \boldsymbol{\theta}_h) \sim \frac{f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h) g(\hat{Y}_{hi} \mid Y_{hi})}{\int f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h) g(\hat{Y}_{hi} \mid Y_{hi}) \mathrm{d}Y_{hi}}. \tag{10}$$

   (a) Generate m samples of Y_{hi}, denoted by $Y_{hi}^{*(j)} (j = 1, \cdots, m)$, from some distribution $h(Y_{hi} \mid \hat{Y}_{hi}, \mathbf{X}_{hi})$ that has the same support of $f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h)$.

   (b) The fractional weight assigned to $Y_{hi}^{*(j)}$ in the t-th EM step is

$$w_{hi1(t)}^{*(j)} \propto \frac{f_1(Y_{hi}^{*(j)} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h^{(t)}) g(\hat{Y}_{hi} \mid Y_{hi}^{*(j)})}{h(Y_{hi}^{*(j)} \mid \hat{Y}_{hi}, \mathbf{X}_{hi})} \tag{11}$$

   with $\sum_j w_{hi1(t)}^{*(j)} = 1$.

2. M-step: Update $\boldsymbol{\theta}_h$ by solving the imputed score function for level 1 model

$$\sum_i \sum_j w_{hi1(t)}^{*(j)} S_1(\boldsymbol{\theta}_h; \mathbf{X}_{hi}, Y_{hi}^{*(j)}) = 0 \tag{12}$$

   where $S_1(\boldsymbol{\theta}_h; \mathbf{x}_{hi}, y_{hi}) = \partial \log f_1(y_{hi} \mid \mathbf{x}_{hi}; \boldsymbol{\theta}_h) / \partial \boldsymbol{\theta}_h$ and $w_{hi1(t)}^{*}$ is computed from (11).

In the E-step, the distribution h is called the proposal distribution while the conditional distribution in (10) is called target distribution. One choice of the proposal distribution is the normal distribution in (6). Once $\hat{\boldsymbol{\theta}}_h$ is obtained from the above EM algorithm, we also need to obtain the variance-covariance matrix of $\hat{\boldsymbol{\theta}}_h$. The observed information based on Louis (1982) formula is given by

$$\begin{aligned} I_{obs}(\boldsymbol{\theta}_h) &= E\{I_{com}(\boldsymbol{\theta}_h) \mid X_h, \hat{Y}_h, \boldsymbol{\theta}_h\} \\ &\quad + E\{S(\boldsymbol{\theta}_h)^{\otimes 2} \mid X_h, \hat{Y}_h, \boldsymbol{\theta}_h\} - \Big[E\{S(\boldsymbol{\theta}_h) \mid X_h, \hat{Y}_h, \boldsymbol{\theta}_h\}\Big]^{\otimes 2} \end{aligned}$$

where $S_1(\boldsymbol{\theta}_h)$ is the complete-sample score function of $\boldsymbol{\theta}_h$, defined in (12), and $I_{com}(\boldsymbol{\theta}_h) = -\partial S_1(\boldsymbol{\theta}_h)/\partial \boldsymbol{\theta}_h'$ is the Fisher information matrix of $\boldsymbol{\theta}_h$. The conditional distribution in computing $I_{obs}(\boldsymbol{\theta}_h)$ can be computed using the above fractionally imputed data after the final EM iteration. The variance-covariance matrix of the MLE of $\boldsymbol{\theta}_h$ is obtained by the inverse of $I_{obs}(\hat{\boldsymbol{\theta}}_h)$.

Now, to discuss parameter estimation for level 2 model, note that we can safely assume that the sampling distribution of $\hat{\boldsymbol{\theta}}_h$ is a normal distribution with mean $\boldsymbol{\theta}_h$ and variance $v_h \equiv I_{obs}^{-1}(\hat{\boldsymbol{\theta}}_h)$. Thus, assuming that f_2 in (2) is also normal distribution with mean $\boldsymbol{\theta}_0$ and variance V_0, we can easily compute the posterior distribution of $\boldsymbol{\theta}_h$ given $\hat{\boldsymbol{\theta}}_h$ and $\zeta = (\boldsymbol{\theta}_0, V_0)$. The posterior distribution is

$$\boldsymbol{\theta}_h \mid (\hat{\boldsymbol{\theta}}_h, \zeta) \sim N\left(\boldsymbol{\theta}_h^*, V_h^*\right)$$

with

$$\hat{\boldsymbol{\theta}}_h^* = \left(V_h^{-1} + V_0^{-1}\right)^{-1}\left(V_h^{-1}\hat{\boldsymbol{\theta}}_h + V_0^{-1}\boldsymbol{\theta}_0\right) \tag{13}$$

and

$$V_h^* = \left(V_h^{-1} + V_0^{-1}\right)^{-1}.$$

In the M-step, the parameters in f_2 are updated by by solving the score equations for $\zeta = (\boldsymbol{\theta}_0, V_0)$ with $\boldsymbol{\theta}_h$ replaced by their conditional expectations computed in the E-step.

4. Prediction

Once the parameter estimates are computed from the two-level EM algorithm discussed in Section 3, we can obtain the best predictor of $Y_h = \sum_i Y_{hi}$. Under the model setup in Section 2, given the parameter estimates, the best predictor of Y_{hi} is given by

$$\hat{Y}_{hi,P} = E(Y_{hi} \mid X_{hi}, \hat{Y}_{hi}; \hat{\boldsymbol{\theta}}_h^*), \tag{14}$$

where the conditional expectation is taken with respect to the distribution in (5) and $\hat{\boldsymbol{\theta}}_h^*$ is computed from (13). Because $\hat{\boldsymbol{\theta}}_h^*$ is used in (14) instead of $\hat{\boldsymbol{\theta}}_h$, the state level prediction $\hat{Y}_{h,P} = \sum_i \hat{Y}_{hi,P}$ borrows strength from the sample observations outside the states. If the conditional distribution is a normal distribution in (6), the best predictor of Y_{hi} is

$$\hat{Y}_{hi,P} = \hat{c}_{hi}\hat{Y}_{hi} + (1-\hat{c}_{hi})(\hat{\boldsymbol{\beta}}_h^{*\prime} X_{hi})$$

where $\hat{\boldsymbol{\beta}}_h^*$ is computed from (13) and $\hat{c}_{hi} = \hat{\sigma}_{hi}^2/(v_{hi} + \hat{\sigma}_{hi}^2)$.

We now discuss estimation of mean squared prediction error (MSPE) of $\hat{Y}_{h,P}$. The MSPE of $\hat{Y}_{h,P}$ is defined by

$$MSPE(\hat{Y}_{h,P}) = E\left\{\left(\hat{Y}_{h,P} - Y_h\right)^2\right\}.$$

Define $\tilde{Y}_{h,P} = \sum_i \tilde{Y}_{hi,P}$ and $\tilde{Y}_{hi,P} = E(Y_{hi} \mid X_{hi}, \hat{Y}_{hi}; \boldsymbol{\theta}_h)$ be the predictor of Y_{hi} using the true parameter values. It can be shown (Kacker and Harville, 1984) that the MSPE of $\hat{Y}_{h,P}$ satisfies

$$MSPE(\hat{Y}_{h,P}) = MSPE(\tilde{Y}_{h,P}) + E\{(\tilde{Y}_{h,P} - \hat{Y}_{h,P})^2\}, \tag{15}$$

where

$$\begin{aligned} MSPE(\tilde{Y}_{h,P}) &= E\{(\tilde{Y}_{h,P} - Y_h)^2\} \\ &= E\{\sum_i V(Y_{hi} \mid \hat{Y}_{hi}, X_{hi})\}. \end{aligned}$$

is the posterior variance of Y_h and the second term of (15) is the additional variability due to using the estimated parameters instead of the true parameters in the prediction. The posterior variance can be computed from the conditional distribution in (5) using $\hat{\theta}_h^*$ in (13). The second term of (15) can be computed from a

Taylor linearization method. In the case of the normal models in (7)-(9), the MSPE estimator is computed by

$$V\{\hat{Y}_{h,P} - Y_h\} = \sum_i \hat{c}_{hi} v_{hi} + \{\sum_i \sum_j (1 - \hat{c}_{hi})(1 - \hat{c}_{hj}) q_{hij}\}, \tag{16}$$

where

$$q_{hij} = (1, X_{hi1}, X_{hi2}) \left\{ V_h^{-1} + \Sigma^{-1} \right\}^{-1} (1, X_{hj1}, X_{hj2})'$$

and V_h is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_h$ from level 1 model.

Instead of the Taylor method, one can also use a parametric bootstrap method to estimate the MSPE. The following parametric bootstrap method can be used to estimate MPSE without relying on the decomposition in (15):

1. Using the parameter estimates $\hat{\boldsymbol{\theta}}_h^*$ and $\hat{\boldsymbol{\alpha}}$, generate $Y_{hi}^{*(b)}$ and $\hat{Y}_{hi}^{*(b)}$ from $f_1(Y_{hi} \mid X_{hi}; \hat{\boldsymbol{\theta}}_h^*)$ and $g(\hat{Y}_{hi} \mid Y_{hi}^*; \hat{\boldsymbol{\alpha}})$, respectively.

2. Using the bootstrap sample of $\hat{Y}_{hi}^{*(b)}$, apply the same procedures to compute the best predictor $\hat{Y}_{h,P}^{*(b)}$ of Y_h.

3. The bootstrap estimator of MSPE of $\hat{Y}_{h,P}$ is computed by

$$B^{-1} \sum_{b=1}^{B} \left\{ \hat{Y}_{h,P}^{*(b)} - Y_h^{*(b)} \right\}^2$$

where $Y_h^{*(b)} = \sum_i Y_{hi}^{*(b)}$.

5 Application to NASS project

We now give a summary of the NASS application for improved crop acreage prediction in the state level estimation for major crops. As mentioned before, a crop is regarded as major for one state if the median of the district crop areas is more than 150,000 acres. Let Y_{hi} be the true crop acreage for district (hi), and we have two auxiliary variables for each district (hi). One is the FSA estimate of Y_{hi}, denoted by X_{hi1}, and the other is the CDL estimate of Y_{hi}, denoted by X_{hi2}. The level 1 model for Y_{hi} incorporating $\mathbf{X}_{hi} \equiv (X_{hi1}, X_{hi2})'$ is

$$Y_{hi} = \mathbf{X}_{hi}' \boldsymbol{\beta}_h + e_{hi}, \tag{17}$$

with $e_{hi} \sim N(0, X_{hi1}\sigma_h^2)$. We use

$$\boldsymbol{\beta}_h \sim N(\boldsymbol{\beta}, \Sigma)$$

as the level 2 model and use

$$\hat{Y}_{hi} | Y_{hi} \sim N(Y_{hi}, \hat{V}_{hi}) \tag{18}$$

as the sampling error model for $\hat{Y}_{hi}$. Thus, the specification of level 1 model is more critical.

Table 1 shows the estimation results from the proposed two-level model for the corn in 2011. The proposed model gives reasonable estimates for the crop acres on the state level. Furthermore, as can be seen in Table 1, the estimation variance is much smaller than the one of JAS estimation.

References

Battesse, G.E., Harter, R,M., and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28-36.

Fay, R.E. and Herriot, R.A. (1979). Estimation of income from small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association* **74**, 269-277.

Fuller, W.A. (2009). *Sampling Statistics*, Wiley.

Table 1: Crop acres estimation from the two-level model for corn in 2011. The first column contains the FIPs for different states, the board estimations are shown in the second column, the third and fourth columns demonstrate the estimations and standard errors of JAS, the fifth and sixth columns contain the estimations and standard errors from the proposed two level model, and the seventh column shows the mean of c_{hi}.

State	board	JAS	JAS.sd	Estimation	Estimation.sd	Mean.chi
IL	126	126.34	2.95	125.74	1.18	9.E-04
IN	59	59.08	2.26	58.18	0.87	2.E-04
IA	141	142.13	3.15	142.14	0.95	3.E-04
KS	49	49.35	3.22	48.77	1.27	1.E-04
KY	13.8	15.76	1.62	13.46	0.24	1.E-04
MI	25	27.26	1.64	23.7	0.84	2.E-01
MN	80.75	81.32	2.89	80.75	0.54	1.E-04
MO	33	35.68	2.03	33.92	0.65	4.E-04
NE	98.5	103.69	3.39	98.7	0.64	2.E-04
ND	22.19	21.9	1.87	20.35	0.8	2.E-01
OH	34	37.27	1.79	35.13	1.76	7.E-01
SD	52	56.58	2.8	52.09	1.97	3.E-01
WI	41.5	42.86	2.39	38.77	0.53	1.E-04

Jiang, J., Lahiri, P., and Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with M-estimation, *The Annals of Statistics*, **30**, 1782–1810.

Kacker, R. and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* **79**, 853-862.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis, *Biometrika*, **98**, 119–132.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B **44**, 226-33.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *International Statistical Review* **70**, 125-144.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators, *Journal of the American Statistical Association* **85**, 163-171.

Rao, J.N.K. (2003). *Small Area Estimation.* Wiley Series in Survey Methodology.

Torabi, M. and Rao, J.N.K. (2008). Small area estimation under a two-level model, *Survey Methodology*, **34**, 11-17.

You, Y. and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, **30**, 431-439.