GENERALIZED ADDITIVE MIXED MODELS IN SMALL AREA ESTIMATION

[1,2]Anang Kurnia, [1]Khairil A. Notodiputro, [1]Asep Saefuddin, [3]I Wayan Mangku

[1]Department of Statistics
Institut Pertanian Bogor, Indonesia

[3]Department of Mathematics
Institut Pertanian Bogor, Indonesia

[2]e-mail : anangk@ipb.ac.id

Abstract. *Small Area Estimation (SAE) is a statistical technique to estimate parameters of sub-population containing small size of samples with adequate precision. This technique is very important to be developed due to the increasing needs of statistic for small domains, such as districts or villages. Some SAE techniques have been developed in Canada, USA, and UE based on real data. We adapted this technique to produce small area statistic in Indonesia based on national data collected by the Statistics Indonesia (Badan Pusat Statistik). We found that the linear model applied to auxiliary data produced estimates with low precision. In this paper we propose a class of generalized additive mixed model to improve the model of auxiliary data in small area estimation.*

Keywords: small area estimation, generalized additive mixed model

1. Introduction

Small Area Estimation (SAE) is the most important concept in survey sampling especially for indirect parameter estimation of relatively small samples. This method can be used to estimate parameters of sub population (a domain which is smaller than population). Direct estimation for sub population fails to provide enough precision because the sample size to yield the estimator is small.

Another method which can be used to obtain higher precision in small area estimation may be developed by linking some information in particular area with some other areas through appropriate model. This procedure is called indirect estimation. The procedure involves data from other domains. In other words, small area estimation model is borrowing strength from sample observation of related areas through auxiliary data (recent census and current administrative records) to increase effective sample size (Rao, 2003).

In this paper we will discuss small area estimation through indirect method or estimation based models. One of the problems found in using this procedure is low precision of linear model for modeling of auxiliary data. In this paper we propose a class of generalized additive mixed model to improve the model of auxiliary data in small area estimation. This paper also presents application on small area estimation using poverty data from Susenas 2005 and Podes 2005 at Bogor District in West Java.

2. Brief Review of Related Topics

Small Area Estimation Based on Linear Mixed Model

There are essentially two-types of models in small area estimation. The first is area level model that relate small area direct estimator to area-specific auxiliary data $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$. We assume the parameter of interest $\theta_i = \mathbf{x}_i'\beta + \upsilon_i$ where $\upsilon_i \sim N(0, A)$ and direct estimator $\hat{\theta}_i = \theta_i + e_i$ where $e_i|\theta_i \sim N(0, D_i)$ and D_i known. The model combines the parameter of interest and the indirect estimates forms $\hat{\theta}_i = \mathbf{x}_i'\beta + \upsilon_i + e_i$ which is a case of generalized linear mixed model. The second is unit level model. In this model the information is available at the sampling unit level and modeling is done based on individual data $\mathbf{x}_{ij} = (x_{1ij}, x_{2ij}, \ldots, x_{pij})$ and we have model $y_{ij} = \mathbf{x}_{ij}^T\beta + \upsilon_i + e_i$ that is a more complex model.

We consider the following Fay-Herriot model (see Fay and Herriot, 1979) for the basic area level model

$$y_i = x_i'\beta + \upsilon_i + e_i$$

where υ_i and e_i are independent with $\upsilon_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$ for $i = 1, 2, \ldots, k$. We assume that β and A unknown but D_i ($i = 1, 2, \ldots, k$) are known.

The best predictor (BP) of $\theta_i = x_i'\beta + \upsilon_i$ if β and A known is given by

$$\hat{\theta}_i^{BP} = \hat{\theta}_i(y_i|\beta, A) = x_i'\beta + (1 - B_i)(y_i - x_i'\beta)$$

where $B_i = D_i/(A + D_i)$ for $i = 1, 2, \ldots, k$. Let $\hat{\theta}_i^{BP} = \hat{\theta}_i(y_i|\beta, A)$ is also Bayes estimator of θ_i under the following Bayesian models:

(i) $y_i|\theta_i \sim N(\theta_i, D_i)$

(ii) $\theta_i \sim N(x_i'\beta, A)$ is prior distribution for θ_i, $i = 1, 2, \ldots, k$.

The Bayes estimator is given from the posterior distribution

$$(\theta_i|y_i, \beta, A) \sim N\left(\left(\frac{y_i}{D_i} + \frac{x_i'\beta}{A}\right)\left(\frac{1}{D_i} + \frac{1}{A}\right)^{-1}\right) = N\left(x_i'\beta + \frac{A}{A+D_i}(y_i - x_i'\beta), \frac{AD_i}{A+D_i}\right)$$

Based on the formulation, we could proof that

$$\hat{\theta}_i^{EB} = E(\theta_i|y_i, \beta, A) = x_i'\beta + (1 - B_i)(y_i - x_i'\beta)$$

where $MSE(\hat{\theta}_i^{EB}) = Var(\theta_i|y_i, \beta, A) = \frac{AD_i}{A+D_i} = (1 - B_i)D_i = g_{1i}(A)$. The estimator $\hat{\theta}_i^{BP}$ are equivalent with $\hat{\theta}_i^{EB}$ for cases that are normally distributed.

When A is known, β could be estimated using the weighted maximum likelihood method

$$\log L(\beta, V) = -\frac{1}{2}\log|V| - \frac{1}{2}(Y - X\beta)' V^{-1}(Y - X\beta)$$

where $V = Diag(A + D_1, A + D_2, \ldots, A + D_k)$.
Let $\beta^* = \hat{\beta}_i(A) = (X'V^{-1}X)^{-1} X'V^{-1}Y$ and by replacing β with β^* in the $\hat{\theta}_i^{BP}$, we get the best linear unbiased predictor (BLUP) of θ_i given by

$$\hat{\theta}_i^{BLUP} = \hat{\theta}_i(y_i|A) = x_i'\beta^* + (1 - B_i)(y_i - x_i'\beta^*)$$

Ghosh and Rao (1994) describe the $MSE(\hat{\theta}_i^{BLUP}) = g_{1i}(A) + g_{2i}(A)$, where

$$g_{1i}(A) = \frac{AD_i}{A+D_i} = (1 - B_i)D_i, \text{ and}$$

$$g_{2i}(A) = D_i^2/(A + D_i)\,[x_i'(X'V^{-1}X)^{-1} x_i]$$
$$= D_i(1 - B_i)\,[x_i'(X'V^{-1}X)^{-1} x_i] \text{ untuk } i = 1, 2, \ldots, k.$$

However, in practice both β and A are unknown. To estimate A, we can use maximum likelihood (ML), restricted/residual maximum likelihood (REML) or method of moment (MM). If we replace β by $\hat{\beta}$ and A by $\hat{A}$ in the BLUP ($\hat{\theta}_i^{BLUP}$) estimator, we get the empirical best linear unbiased predictor (EBLUP)

$$\hat{\theta}_i^{EBLUP} = \hat{\theta}_i(y_i|\hat{A}) = x_i'\hat{\beta} + (1 - \hat{B}_i)(y_i - x_i'\hat{\beta})$$

If defined MSE of $\hat{\theta}_i^{EBLUP}$ is $MSE(\hat{\theta}_i^{EBLUP}) = E(\hat{\theta}_i^{EBLUP} - \theta_i)^2 = Var(\hat{\theta}_i^{EBLUP}) + (Bias\ \hat{\theta}_i^{EBLUP})^2$, Kacker and Harville (1984) reformulated it as

$$\begin{aligned} MSE(\hat{\theta}_i^{EBLUP}) &= MSE(\hat{\theta}_i^{BLUP}) + E(\hat{\theta}_i^{EBLUP} - \hat{\theta}_i^{BLUP})^2 \\ &= H_{1i}(A) + H_{2i}(A) \end{aligned}$$

where $H_{1i}(A) = MSE(\hat{\theta}_i^{BLUP}) = g_{1i}(A) + g_{2i}(A)$ and $H_{2i}(A) = E(\hat{\theta}_i^{EBLUP} - \hat{\theta}_i^{BLUP})^2$. Leading term $g_{1i}(A)$ lead to large reduction in MSE relative to the MSE of the direct estimator, $g_{2i}(A)$ is due to estimating of β and $H_{2i}(A)$ is due to estimating A.

Prasad and Rao (1990) used the Taylor series method to estimate $g_{1i}(A)$, $g_{2i}(A)$ and $H_{2i}(A)$. The MSE estimator of $\hat{\theta}_i$

$$MSE(\hat{\theta}_i^{EBLUP})^{PR} = g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2\, g_{3i}(\hat{A})$$

where $g_{3i}(\hat{A}) = \dfrac{2D_i^2}{k^2(A+D_i)^3} \sum_{j=1}^{k}(A+D_j)^2$. The $MSE(\hat{\theta}_i)^{PR}$ is identical to the Bayes risk as defined by Butar and Lahiri (2003).

Generalized Additive (Mixed) Model

Multiple regression analysis is one of the most widely used statistical techniques. It is a powerful tool when its assumptions are met, including that the relationships between the predictors and the response are well described with a defined function (e.g., straight-line, polynomial, or exponential). In many applications, however, the reliance on a defined function is limited. Many phenomena do not have a relationship that can be easily defined.

To overcome the above difficulties, Stone (1985) proposed the additive model to solve them. These models estimate an additive approximation to the multivariate regression function. The advantages of this approximation are at least twofold. First, since each of the individual additive terms is estimated using a univariate smoother, the curse of dimensionality is avoided, at the cost of not being able to approximate universally. Second, estimates of the individual terms explain how the dependent variable changes with the corresponding independent variables.

In general, generalized additive models (GAM) enable us to relax this assumption by replacing a defined function with a non-parametric smoother to uncover existing relationships. Smoothing is a method that will highlight a trend by separating it from variability due to noise. Several different smoothers are available, but the most commonly used are spline or loess. Smoothers have a parameter that can be used to control the closeness of the fit of the trend to the data. For detail about GAM, Hastie and Tibshirani (1990).

GAM are additive models because they simultaneously fit the distinct effects of each independent variable. Each effect can be estimated using either a smoother or a defined function, leading to the description of GAM as semiparametric. GAM are appropriate under the assumption of the absence of interaction effects.

GAM also offers the added flexibility of permitting non-normal error distributions. This allows modeling response variables with distributions such as binomial and Poisson. Generalized Additive Mixed Models

(GAMM) have also been recently developed to incorporate random effects, which are an additive extension of Generalized Linear Mixed Model (GLMM) in the spirit of Hastie and Tibshirani (1990).

Let Y be a response random variable and $X_1, X_2, \ldots, X_p$ be a set of predictor variables. A regression procedure can be viewed as a method for estimating the expected value of Y given the values of $X_1, X_2, \ldots, X_p$. The standard linear regression model assumes a linear form for the conditional expectation

$$E(Y \mid X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Given a sample, estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are usually obtained by the least squares method. The additive model generalizes the linear model by modeling the conditional expectation as

$$E(Y \mid X_1, X_2, \ldots, X_p) = s_0 + s_1(X_1) + s_2(X_2) + \ldots + s_p(X_p)$$

where $s_i(X)$, $i = 1,2, \ldots, p$ are smooth functions.

In order to be estimable, the smooth functions s_i have to satisfy standardized conditions such as $Es_i(X_i) = 0$. These functions are not given a parametric form but instead are estimated in a nonparametric fashion. While traditional linear models and additive models can be used in most statistical data analysis, there are types of problems for which they are not appropriate. For example, the normal distribution may not be adequate for modeling discrete responses such as counts or bounded responses such as proportions.

Generalized additive models address these difficulties, extending additive models to many other distributions besides just the normal. Thus, generalized additive models can be applied to a much wider range of data analysis problems. Similar to generalized linear models, generalized additive models consist of a random component, an additive component, and a link function relating the two components. The response Y, the random component, is assumed to have exponential family density. The mean of the response variable μ is related to the set of covariates $X_1, X_2, \ldots, X_p$ by a link function g. The quantity

$$\eta = s_0 + \Sigma s_i(X_i)$$

defines the additive component, where $s_1(\cdot), \ldots, s_p(\cdot)$ are smooth functions, and the relationship between μ and η is defined by $g(\mu) = \eta$. The most commonly used link function is the canonical link, for which $\eta = \theta$.

Furthermore, Lin and Zhang (1999) proposed Generalized Additive Mixed Models (GAMM) for overdispersed and correlated data. They explored the Generalized Linear Mixed Model (GLMM) representation of the smoothing spline estimators and estimated the smoothing parameter using REML. Following Breslow and Clayton (1993), Lin and Zhang (1999) used Double Penalized Quasi-Likelihood to estimate beta and REML is used to estimate the variance components.

3. The GAMM Approach for Small Area Estimation

Rao (2003) gives extensive review of the most commonly used estimators, including synthetic and composite estimator, empirical best unbiased linear predictors, empirical Bayes and hierarchical Bayes approach. All of them in use for small area estimation based on parametric approach. In this chapter we propose a class of nonparametric approach, generalized additive mixed model (GAMM). The GAMM approach has significant advantages over its parametric approach to model auxiliary variable, and then we adopt this approach to application in small area estimation.

We consider an extension of the Fay-Herriot model for the basic area level model

$$y_i = x_i'\beta + \upsilon_i + e_i, \; i = 1, 2, \ldots, k$$

where β is coefficient regression parameters, υ_i are random effect area, and e_i are sampling errors. We also assume $e_i \sim (0, D_i)$, $\upsilon_i \sim (0, A)$ and that they are independent. D_i is usually assumed to be known, see Rao (2003).

We assume that y_i and x_i are related by a smooth function m(.). Let X be the random vector of predictors, thus

$$y_i = m(x_i) + \upsilon_i + e_i\,,\ i = 1, 2, ..., k$$

where $\upsilon_i|X \sim (0, \upsilon(x_i))$, $e_i \sim (0, D_i)$, and e_i and υ_i are independent. The small area mean functions is

$$\theta_i(x_i) = m(x_i) + \upsilon_i$$

are linear combination of mean $m(x_i)$ and the random effects υ_i. We can use an estimator of the mean function using a linear smoother such as smoothing splines, regression splines, and local polynomial regression. For detail discussion of these methods, see Hastie and Tibshirani, (1990).

If we use Kernel smoothing function to estimate $m(x_i)$, the best predictor for small area means θ_i can be written as

$$E(\theta_i|y_i) = \gamma_i\, y_i + (1 - \gamma_i)\, \hat{m}_h(x_i)$$

where $\gamma_i = \upsilon(x_i) / (\upsilon(x_i) + D_i)$. To approximate MSE, we substitute $x_i'\beta$ in linear mixed model with $\hat{m}_h(x_i)$.

$$\text{mse}(\hat{\theta}_i) = \frac{D_i\hat{\sigma}_u^2}{D_i + \hat{\sigma}_u^2} + (1-\hat{\gamma})^2\, \text{mse}\left(\hat{m}_h(x_i)\right) + 2D_i^2\left(\hat{\sigma}_u^2 + D_i\right)^{-3}\, \text{mse}\left(\hat{\sigma}_u^2\right)$$

4. Empirical Application and Discussion

Our empirical studies used two data set. The first, was hypothetic data for 32 small area where υ_i and e_i have normal distribution with mean 0 and variance 1. Y, which is the variable that we are interested is, define as function of X^2 and X is auxiliary data. GAMM approach show better prediction than EBLUP estimator. The mean absolute relative estimation (MARE) of GAMM approach is 0.0193 and the EBLUP estimator is 0.0212. Further, the relative root mean square error (RRMSE) of GAMM approach is 0.0289, while the EBLUP estimator is 0.0327

The second data set, was real data for PODES 2005 and SUSENAS 2005 especially for Bogor Municipality. Both data were collected by BPS (Statistics Indonesia). Y is unemployment level which is indicated as percentage of unemployment from group of "age work" for each village in Bogor Municipality. Percentage of men (X2), percentage of non-permanent housing (X5), percentage of letter poor statement (X7), and percentage of pre prosperous-family and prosperous-family 1 (X8) are used as auxiliary variable.

Table 1. Estimator of Unemployment Level in Bogor Municipality

	Village	Direct	GAMM	EBLUP		Village	Direct	GAMM	EBLUP
1002	Pamoyanan	13.04	12.64	13.03	4006	Sempur	10.94	10.38	10.93
1005	Kertamaya	8.42	8.86	8.43	4010	Kebonkelapa	12.07	12.06	12.07
1006	Rancamaya	25.00	23.36	24.94	5002	Pasirkuda	20.00	17.60	19.95
1009	Muarasari	1.85	1.97	1.85	5003	Pasirjaya	13.51	12.91	13.49
1013	Batutulis	6.38	6.46	6.39	5004	Gunungbatu	10.64	10.31	10.63
1015	Empang	3.33	3.42	3.34	5006	Menteng	10.91	10.91	10.90
1016	Cikaret	9.80	9.74	9.80	5008	Cilendek Barat	16.67	15.81	16.64
2002	Sindangrasa	1.67	1.75	1.67	5009	Sindangbarang	6.38	6.72	6.39
2006	Sukasari	8.33	8.21	8.33	5012	Situgede	4.00	4.24	4.00
3001	Bantarjati	5.45	5.56	5.46	5015	Curugmekar	10.42	10.25	10.41
3002	Tegalgundil	6.90	6.98	6.90	6001	Kedungwaringin	6.38	6.33	6.39
3004	Cimahpar	3.28	3.59	3.29	6003	Kebonpedes	9.43	9.55	9.44
3006	Cibuluh	10.53	10.91	10.53	6004	Tanahsareal	11.54	10.92	11.53
3007	Kedunghalang	9.09	8.94	9.09	6005	Kedungbadak	6.38	6.35	6.38
3008	Ciparigi	4.88	5.16	4.88	6007	Sukadamai	12.50	11.99	12.49
4002	Gudang	14.81	14.48	14.79	6009	Kayumanis	5.45	5.56	5.47
4004	Tegallega	2.27	2.53	2.28	6011	Kencana	6.25	6.57	6.26

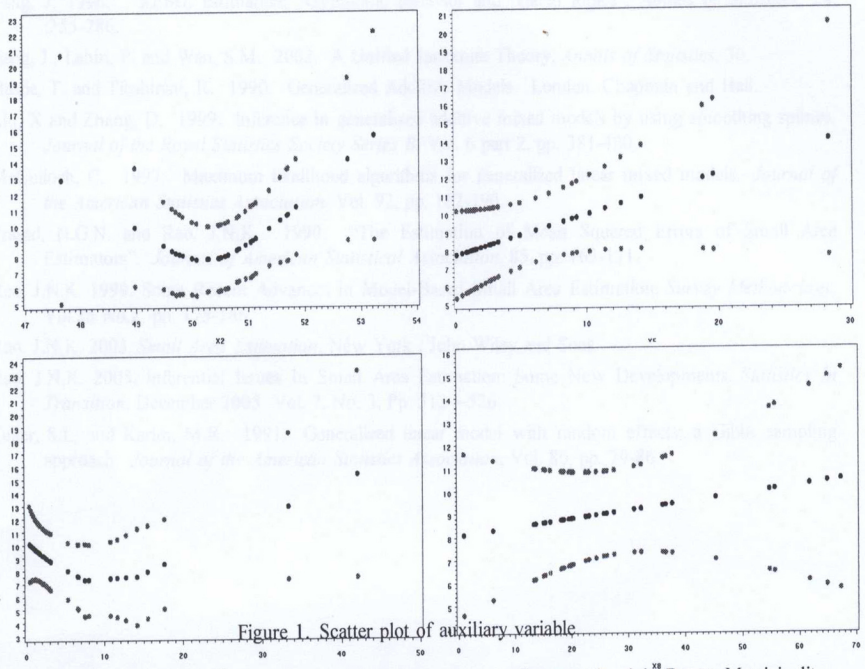

Figure 1. Scatter plot of auxiliary variable

Table 1 exhibits the results from each method to estimate unemployment level in Bogor Municipality. The RRMSE for direct estimator, GAMM approach and EBLUP are 0.0361, 0.0326 and 0.0335. Actually all of the estimators support direct estimator. The possible factors which can affect this condition is variance between small area that was higher than variance sampling error within small area. However, the GAMM approach was able to reduce the auxiliary variable influence which was not linear. Figure 1 shows the scatter plot of auxiliary variable while X2 and X7 have not linearity between the auxiliary and the response interest.

It is shown in our study that generalized additive mixed model outperforms generalized linear mixed model in EBLUP at least in two aspects. First, generalized additive mixed model relaxes the assumption of linearity between the predictors and the response and avoids the problem of model misspecification that often happened in EBLUP. Secondly, by incorporating nonlinear effects, generalized additive mixed model helps to discover the hidden pattern of predictors and therefore improves the predictive performance.

REFERENCES

Butar, F. B. and Lahiri, P. 2003. On measures of uncertainty of empirical Bayes small area estimators. Model selection, model diagnostics, empirical Bayes and hierarchical Bayes, *Journal of Statistical Planning and Inference*, 112, pp: :63–76.

Breslow, N.E. and Clayton, D.G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistics Association*, Vol. 88, pp. 9-25.

Fahrmeir, L. and Lang, S. 2001. Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Applied Statistics*, Vol. 50, part 2, pp. 201-220.

Fay, R.E. and Herriot, R.A., (1979), "Estimates of income for small places: An application of James-Stein procedures to Census data". *Journal of the American Statistical Association*, Vol. 74, p:269-277

Ghosh, M. and Rao, J.N.K. 1994. "Small Area Estimation: An Appraisal". *Statistical Science*, 9, No.1 p:55-93.

Jiang, J. 1996. "REML estimation: Asymptotic behavior and related topics", Annals *of Statistics*, 24, :255-286.

Jiang, J., Lahiri, P. and Wan, S.M. 2002. A Unified Jackknife Theory, *Annals of Statistics*, 30.

Hastie, T. and Tibshirani, R. 1990. Generalized Additive Models. London: Chapman and Hall.

Lin, X and Zhang, D. 1999. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistics Society Series B*, Vol. 6 part 2, pp. 381-400.

McCulloch, C. 1997. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistics Association*, Vol. 92, pp. 162-190.

Prasad, N.G.N. and Rao, J.N.K. 1990. "The Estimation of Mean Squared Errors of Small Area Estimators". *Journal of American Statistical Association*, 85, pp. 163-171.

Rao, J.N.K. 1999. Some Recent Advances in Model-Based Small Area Estimation, *Survey Methodology*, Vol.25 No.2, pp. 175-186.

Rao, J.N.K. 2003. *Small Area Estimation*, New York : John Wiley and Sons.

Rao, J.N.K. 2005. Inferential Issues In Small Area Estimation: Some New Developments. *Statistics In Transition*, December 2005 Vol. 7, No. 3, Pp. 513—526.

Zeger, S.L. and Karim, M.R. 1991. Generalized linear model with random effects: a Gibbs sampling approach. *Journal of the American Statistics Association*, Vol. 86, pp. 79-86.