Small Area Modeling for Survey Data with Smoothed Error Covariance Structure via Generalized Design Effects

A.C. Singh[1], R.E. Folsom, Jr[2]., and A.K. Vaish[2]
Statistics Canada[1], 16-Q, R.H. Coats, 120 Parkdale Ave, Ottawa, ON K1A 0T6, avi.singh@statcan.ca
RTI International[2] , 3040 Cornwallis Road, PO Box 12194, Durham, NC 27709, ref@rti.org, avaish@rti.org

Abstract

We consider specifying the design-based error covariance structure in small-area modeling with survey data. While it is customary to treat the estimated covariance as known, it is often unstable. To alleviate this problem, one can either model the distribution of design-based covariance matrix or smooth the estimated covariance by specifying only its mean function. We prefer smoothing over modeling because of the strong assumptions required to model the distribution of the error covariance structure for small area estimates (SAEs). To smooth nondiagonal error covariance matrices, we make use of the g-deff (generalized design effect), defined earlier by Rao and Scott (1981) in the context of categorical data analysis. Simulation results for SAEs based on a linear mixed model show that the proposed smoothing provides improved coverage of confidence intervals.

Key Words: Estimating functions, Generalized design effects, Ignorable and nonignorable designs; Unstable estimated error covariance matrix

1. Introduction

In any modeling problem, one needs to specify the mean function and the error covariance structure under a semi-parametric approach based on the first two moments. In the case of modeling direct domain total estimates $\{ t_{y,d} : d = 1,..., D \}$ of small areas (see e.g., Fay and Herriot, 1979) from surveys, special problems arise in specifying the error covariance structure V_t. However, since an approximately unbiased estimate $\hat{V}_t$ of V_t may be available, it is customary (see Rao, 2003, pp. 76) for such models to treat $\hat{V}_t$ as known. For large samples, such an assumption is, of course, commonly made. However, for small samples n_d , this is clearly not desirable because like the direct point estimates $t_{y,d}$, the direct variance estimates $\hat{V}_{t(d)}$ are also subject to instability, and so treating $\hat{V}_{t(d)}$ as known may cause serious underestimation of variance of SAEs. As an alternative, one may want to model $\hat{V}_t$ in addition to $t_{y,d}$ which renders the small area estimation (SAE) problem even more complicated as the variance of $\hat{V}_t$ involves unknown third and fourth order sample inclusion probabilities. Under certain simplifying assumptions about the sample design, and the superpopulation model, the problem can, however, be simplified; see e.g., the use of a Wishart distribution for $\hat{V}_t$ by Otto and Bell (1995). Nevertheless, in practice, it would be desirable to make rather weak assumptions about $\hat{V}_t$ that seem plausible, because practitioners, in general, prefer to take the path of least assumptions in modeling.

In trying to model $\hat{V}_{t(d)}$ under weak assumptions, it may be useful to first observe that the problem of modeling the mean of $\hat{V}_t$ is much simpler than specifying the full distribution of the error covariance structure, i.e., modeling the variance of $\hat{V}_t$. In this paper we suggest that smoothing $\hat{V}_t$ might provide a reasonable practical compromise between the two extremes of modeling or no modeling of $\hat{V}_t$.

To smooth the $\hat{V}_t$ matrix, we propose to use the generalized design effect (g-deff) considered earlier by Rao and Scott (1981) in the context of categorical analysis of survey data. G-deffs are defined as the eigen-values (λ_t's) of the matrix product $V_t^{*-1} V_t$ where V_t^* is the error covariance under a suitable working assumption such as that of the simple random sampling design or that the design is ignorable for the superpopulation model under consideration.

In the case of the unit-level superpopulation model, the problem of smoothing $\hat{V}_t$ becomes considerably more involved than for the aggregate-level (Fay-Herriot) case because of additional summary statistics that enter in the picture from the theory of EFs. Note that Singh, Folsom, and Vaish (2002, 2003) proposed an estimating function-based Gaussian likelihood (EFGL) methodology in a hierarchical Bayes framework to generalize the usual SAE modeling of Fay and Herriot (1979) at the aggregate level to unit level modeling. By using more detailed information, the unit-level model typically yields efficiency gains relative to the aggregate model. However, they didn't specify the summary statistics explicitly. In this paper, the vector ψ of optimal summary statistics is specified explicitly which makes it computationally easier to use g-deff for smoothing the error covariance structure $\hat{V}_\psi$ under unit-level modeling.

In Section 2, we consider first the general problem of modeling with survey data, and then the choice of appropriate finite population parameters and the corresponding sample summary statistics in the context of SAE modeling. We consider both aggregate and unit level modeling and identify the problem of instability of $\hat{V}_\psi$ in each case for the corresponding vector ψ . In the next section 3, we review application of the usual deff in generalized variance function modeling, and its use for smoothing $\hat{V}_t$ when it is diagonal. We next describe the proposed method of g-deff for

smoothing nondiagonal $\hat{V}_t$ or $\hat{V}_\psi$. In Section 4, we present a simulation study for the linear mixed superpopulation model and show how EFGL-smoothed compares with EFGL-unsmoothed. In the simulation study, the case of aggregate level modeling was, however, not considered although it follows readily from the general case by replacing the unit-level covariate value x_{dk} with the domain level average value $A_{x,d}$ for each unit k in domain d. Finally, Section 5 contains the summary and some concluding remarks.

2. Modeling with Survey Data

The difficulty in modeling survey data is well known in view of the two basic results of Godambe (1955, 1966): first, the nonexistence of a uniformly minimum variance unbiased estimate in a suitable linear class which causes difficulty in using a semiparametric approach, and second, likelihood being flat for the unseen (i.e., nonselected population units) given the seen and thus making it difficult to use the likelihood approach either in a frequentist or a Bayesian framework. The main reason underlying these problems that distinguish survey modeling from mainstream statistics is that there are too many finite population parameters to cope with if one identifies each unit's characteristic as a parameter of interest. The reality is that in practice we are not interested in characteristics at the unit level, but instead we need to define suitable finite population quantities corresponding to a group of units. For the difficult but realistic problem of nonignorable designs for a given superpopulation model, this can be done using census EFs (i.e., assuming the sample is the census, see, e.g., Binder, 1983) which depend, of course, on the model parameters to be estimated. For the unit-level model the estimating functions can be specified as follows:

$$\phi_{\eta(d)} = \sum_k (y_{dk} - x'_{dk}\beta - \eta_d) w_{dk} = t_{y,d} - t'_{x,d}\beta - t_{c,d}\eta_d$$
$$\phi_\beta = \sum_d \sum_k x_{dk}(y_{dk} - x'_{dk}\beta - \eta_d) w_{dk} = t_{xy} - t'_{xx'}\beta + \sum_d t_{x,d}\eta_d \qquad (2.1)$$

The optimal summary statistics ψ for estimating (β, η_d) are obtained as $\{t_{y,d}, t_{x,d}, t_{c,d}, t_{xy}, t_{xx'}\}$ with somewhat self-explanatory new notations for certain estimated population totals. In this case, the unit-level model for this set of summary statistics can be expressed with the set of incidental parameters $\{p_d, \mu_{x,d}, \mu_{xx'}\}$ as :

$$\begin{aligned} t_{y,d} &\approx N p_d \mu_{y,d} + e_{y,d} \\ t_{c,d} &\approx N p_d + e_{c,d} \\ t_{x,d} &\approx N p_d \mu_{x,d} + e_{x,d} \\ t_{xy} &\approx N(\mu_{xx'}\beta + \textstyle\sum_d p_d \mu_{x,d}\eta_d) + e_{xy,d} \\ t_{xx'} &= N\mu_{xx'} + e_{xx'} \end{aligned} \qquad (2.2)$$

where $\mu_{y,d}$ is the limit of $t_{y,d}/N_d$ as $N_d \to \infty$, $\mu_{x,d}$ is similarly defined, and p_d is the limit of N_d/N. While the model (2.2) is now nonlinear because of the incidental parameters it turns out to be linear for computing conditional posteriors, and so no new computational complexity is involved except that there are more steps in the MCMC cycles because of extra parameters. It may be noted that for unit-level modeling, the covariates x_{dk}'s are typically taken as categorical (such as demographic group indicators) because the domain totals $T_{x,d}$ or averages $A_{x,d}$ required for each variable may only be available in practice as domain counts or proportions for each covariate category.

3. Proposed Method of Smoothing $\hat{V}_\psi$ Using g-deff

First we review the use of Deff for generalized variance function (GVF) modeling.

3.1 Deff for GVF modeling

Suppose V_ψ is a $D \times D$ diagonal matrix, and let γ_d be the design effect for the domain-d SAE, i.e., $\gamma_d = V_{\psi(d)} / V^*_{\psi(d)}$. In view of the fact that the design effects γ_d are often approximately constant over a suitable set of statistics, a simple type of GVF modeling assumes that the mean of $\hat{V}_{\psi(d)}$ is proportional to $V^*_{\psi(d)}$ where the covariate $V^*_{\psi(d)}$ is taken to be approximately known. As mentioned earlier, $V^*_{\psi(d)}$ is estimated under the assumption of simple random sampling or that the design is ignorable for the model and that the estimate $\hat{V}^*_{\psi(d)}$ is assumed to be stable. So for the GVF model with $d = 1, ..., D$,

$$\hat{V}_{\psi(d)} = \gamma V^*_{\psi(d)} + e_{v,d} \qquad (3.1)$$

where the error covariance structure is not specified, the mean parameter γ can be estimated by the average of $\hat{\gamma}_d$ over domains where $\hat{\gamma}_d = \hat{V}_{\psi(d)} / \hat{V}^*_{\psi(d)}$. The smoothed estimate of $V_{\psi(d)}$ is then obtained as $\tilde{V}_{\psi(d)} = \hat{\gamma}\hat{V}^*_{\psi(d)}$. If it does not seem reasonable to assume that the deff is constant over all areas, then one can compute separate estimates of γ for subgroups of areas for which the constant deff assumption seems plausible. Note that $V^*_{\psi(d)}$ may depend on the mean parameters $\{\beta, \eta_d\}$ of the small area model as in the case of binary data, and then the smoothed variance estimate can also be allowed to depend on unknown mean parameters. When dealing with discrete data, this is clearly a desirable feature.

3.2 GVF-type Modeling for Non-diagonal $\hat{V}_\psi$

The above GVF modeling to smooth $\hat{V}_\psi$ is not applicable when $\hat{V}_\psi$ is nondiagonal because the concept of deff is not defined for off-diagonal terms of covariances. In this case, using the concept of g-deff (see e.g., Rao and Scott, 1981) defined as eigen-values ($\lambda_k, k = 1, ..., K$) of $V^{*-1}_\psi V_\psi$, K being

the dimension of vector ψ, we can write a GVF-type model as

$$vec(\hat{V}_\psi) = \sum_j \lambda_j \, vec(\sum_k q_{jk} q'_{jk}) + vec(e_v) \quad (3.2)$$

where 'vec' notation is used to signify that columns of the matrix are stacked one above the other, the second sum is over the j^{th} subgroup of summary statistics, and the first sum is over all subgroups. The above model is motivated by the matrix result for a pair of real symmetric matrices with at least one of them being positive definite (cf: C.R. Rao, 1973, pp.41) which states that there exists a nonsingular matrix Q such that (here the pair of matrices are V_ψ and V_ψ^*),

$$\begin{aligned} V_\psi &= Q\Lambda Q' = \sum_k^K \lambda_k q_k q_k' \\ V_\psi^* &= QQ' = \sum_k^K q_k q_k' \end{aligned} \quad (3.3)$$

where $\Lambda = diag(\lambda_k)$, and q_k is the k^{th} column of Q. It follows from (3.3) that if $\hat{V}_\psi^*$ as well as $\sum_k^{K_j} q_k q_k'$ for each selected subgroup of areas are stable, then the instability of $\hat{V}_\psi$ can be overcome by smoothing the estimated eigen-values $\hat{\lambda}_k$ over subgroups. Note that the eigen-values are nonnegative for a nonnegative definite real symmetric matrix. Thus as with deff, if the estimated g-deffs are averaged over suitable subgroups ($j = 1,...,J$) as defined by the GVF-type model (3.2), the smoothed estimate of V_ψ is obtained as

$$\tilde{V}_\psi = \sum_j \hat{\lambda}_j \, (\sum_k q_{jk} q'_{jk}) \quad (3.4)$$

3.3 Aggregate-level vs. Unit-level Modeling

Finding eigen-values and eigen-vectors for g-deff based smoothing could be computationally difficult if the dimension K is large. For the aggregate-level model, $\hat{V}_\psi$ is typically block-diagonal, and so the proposed method of g-deff is not too complicated computationally. However, for unit-level modeling, it remains nondiagonal because of summary statistics (obtained from EFs for fixed parameters) that aggregate over all areas. In these situations, one can take advantage of certain patterns that typically arise in $\hat{V}_\psi$. Observe that the small areas or domains can often be grouped in practice into strata or superstrata, and then $\hat{V}_\psi$ can be partitioned as

$$\hat{V}_\psi = \begin{pmatrix} A & B \\ B' & C \end{pmatrix} \quad (3.5)$$

where A is a high dimensional block diagonal matrix corresponding to domain or strata-level summary statistics, and C is a low dimensional matrix corresponding to summary statistics aggregated over all domains. Now decompose $\hat{V}_\psi$ as

$$\begin{pmatrix} A & B \\ B' & C \end{pmatrix} = H^{-1} \begin{pmatrix} A & O \\ O & C - B'A^{-1}B \end{pmatrix} H^{-1'} \text{ where} \quad (3.6)$$

$$H = \begin{pmatrix} I & O \\ -B'A^{-1} & I \end{pmatrix}.$$

The above matrix $C - B'A^{-1}B$ is expected to be stable and the effect of instability of the matrix H on $\hat{V}_\psi$ is expected to be subsumed in that of the matrix A because covariance of the transformed vector $H\psi$ turns out to be block-diag $\{A, C - B'A^{-1}B\}$. So it is probably sufficient to just smooth A to obtain $\tilde{A}$ (using the g-deff idea) which being block-diagonal is not computationally demanding, and then obtain $\tilde{V}_\psi$ from the decomposition in (3.8) wherein the matrices H and $C - B'A^{-1}B$ are not modified or smoothed.

4. Simulation Study

We design our study along the lines of Singh, Folsom, and Vaish (2003) which is based on Pfeffermann et al. (1998). Consider a universe of $d = 1, \cdots, D$ strata (small areas) where $D = 100$ and let N_d denote the number of population members in stratum-d. In this simulation experiment, we set $N_d = N_0 \, (1 + \exp(u_d^*))$ where N_0 is a constant and u_d^* is obtained by truncating $u_d \sim N(0, 0.2)$ at $\pm\sqrt{0.2}$. For simplicity, we consider a single covariate super-population linear mixed model $y_{dk} = \beta_0 + x_{dk}\,\beta_1 + \eta_d + \varepsilon_{dk}$ where $\beta_0 = 0.5$, $\beta_1 = 1$, $\eta_d \sim N(0, 0.2)$, $\varepsilon_{dk} \sim N(0,1)$, and $k = 1, \cdots, N_d$. The covariate is $x_{dk} = \upsilon_d + \delta_{dk}$ where $\upsilon_d \sim N(0, 0.1)$ and $\delta_{dk} \sim N(0,1)$. We generate $M = 500$ population level data sets with common x_{dk} and N_d where N_d's are generated using N_0 =3000. We selected a sample from each of these populations so that the design was nonignorable. To select a sample with nonignorable design, we stratify the stratum-d population into two substrata Ω_{d+} with $\varepsilon_{dk} > 0$ and Ω_{d-} with $\varepsilon_{dk} \le 0$. Let N_{d+}, N_{d-} denote the sizes of these substrata and n_{d+}, n_{d-} denote the sizes of the simple random samples selected without replacement from these strata, respectively. Note that the substratum sizes vary across the 500 populations. We have $N = \sum_{d=1}^{100} N_d$ and $n = \sum_{d=1}^{100} n_d$ where $n_d = n_{d-} + n_{d+}$. For 500 populations, we generate the corresponding 500 samples. In our simulation experiment, $(n_{d+}, n_{d-}) = (5,15)$, so that we have common sample sizes from each area.

The results in this paper for the EFGL-unsmoothed method are not directly comparable to those in the Singh, Folsom, and Vaish (2003) paper because here we are using the version of EFGL based on the summary statistics ψ which consists of $\{t_{y,d}, t_{x,d}, t_{xy}, t_{x^2}\}$; $t_{c,d}$ is excluded because it is constant and

equal to N_d for the stratified SRS design. Note that the simpler case of aggregate-level modeling was not considered in this limited simulation study. To avoid computational complexity, we used the smoothing based on the decomposition (3.6) where A is a block diagonal matrix of 100 blocks, each of dimension 2x2 corresponding to each small area, and $C - B'A^{-1}B$ is only a 2x2 matrix. In order to use EFGL under a HB framework, customary priors for β and σ_η^2 were chosen. Estimation of the parameter σ_ε^2 was not considered as it was not needed.

The results from the simulation study are presented in Tables 1, 2, and 3. In terms of the model parameter posterior means and standard deviations, both methods (EFGL-u for unsmoothed, and EFGL-s for smoothed) perform very similarly as seen from Table 1. However, in terms of coverage probabilities the unsmoothed method EFGL-u does not perform as well as EFGL-s as seen from Table 2. Table 3 shows that the median coefficient of variation of the estimated mean squared errors over 100 small areas was ~48% for EFGL-u vs. ~38% for EFGL-s. This improved stability in the estimated SAE mean squared errors for the EFGL-s solution accounts for its superior coverage probabilities.

5. Summary and Concluding Remarks

In this paper we considered the problem of smoothing the error covariance matrix in small area modeling of survey data. This problem has been around for quite some time. The proposed method of smoothing the error covariance based on deff and g-deff provides a simpler alternative to other methods including modeling the distribution of the error covariance. The g-deff smoothing seemed to perform well in a limited simulation study. An important consideration in using the proposed method is that similar to GVF modeling, the underlying assumptions are quite mild; this is likely to be attractive to practitioners. An interesting finding of the simulation study was that although smoothing seemed to help cure instability of variance of SAEs, it may not be sufficient for areas with very small sample sizes. In future, it would be interesting to see if suitable collapsing of areas would help to provide a sufficiently stable smoothed estimate of error covariance as well as help in justifying the normal approximation for summary statistics. We also note that the basic idea of covariance smoothing proposed in this paper is applicable to other SAE problems as well involving spatial and temporal modeling.

Acknowledgments: The first author's research was supported in part by a grant from Natural Sciences and Engineering Research Council of Canada held at Carleton University, Ottawa under an adjunct research professorship.

References

Binder, D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review,* **51**, pp. 279-292.

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*, **74**, pp. 269-277.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *JRSS (B)*, 17, 269-278.

Godambe, V.P. (1966). A new approach to sampling from finite populations I, II (with discussion), *JRSS (B)*, 28, 310-328.

Otto, M.C., and Bell, W.R. (1995), "Sampling Error Modeling of Poverty and Income Statistics for States," *Proceedings of the Government Statistics Section, American Statistical Association*, pp. 160-165.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J. (1998), "Weighting for Unequal Selection Probabilities in Multilevel Models," *Journal of the Royal Statistical Society*, B, **60**, pp. 23-40.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, Second Ed., New York: John Wiley.

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

Rao, J.N.K., and Scott, A.J. (1981) The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness-of-fit and independence in two-way tables, *Journal of the American Statistical Association,* 76 221-230.

Singh, A.C., Folsom, R.E., Jr., and Vaish, A.K. (2002). Estimating function-based approach to hierarchical Bayes Small Area Estimation from survey data, I*nternational Conference on Recent Advances in Survey Sampling,* , in honour of J.N.K. Rao's 65th birthday, July 10-13.

Singh, A.C., Folsom, R.E., Jr., and Vaish, A.K. (2003). Hierarchical Bayes small area estimation for survey data by EFGL: the method of estimating function-based Gaussian Likelihood (with discussion), *FCSM Statistical Policy Working paper* #36, pp. 47-74 (www.fcsm.gov/working-papers/).

Table 1. Average Posterior Mean and Standard Deviation (SD) for Model Parameters

Parameter (true value)	Average Posterior Mean		Average Posterior SD	
	EFGL-u	EFGL-s	EFGL-u	EFGL-s
$\beta_0(0.5)$	0.4944	0.4949	0.0475	0.0475
$\beta_1(1.0)$	0.9996	0.9990	0.1228	0.1229
$\sigma_\eta^2(0.2)$	0.1971	0.1966	0.0317	0.0317

Table 2. 95% Normal Coverage Probability and Ratio of Prediction (PI) Widths

Percentiles and Mean Over Small Areas	Coverage Probability		Ratio of Average PI Widths
	EFGL-u	EFGL-s	EFGL-s/EFGL-u
75%	0.9240	0.9460	1.02
Mean	0.9186	0.9396	1.01
25%	0.9120	0.9340	1.00

Table 3. Median of Mean Squared Errors (MSEs) Over 100 Small Areas

Median Over 100 Small Areas	EFGL-u	EFGL-s
Simulated True MSE	0.023	0.023
Estimated MSE Expected Values	0.023	0.023
Coefficient of Variation of Estimated MSEs	48.37%	37.75%