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] and RTI International[2]
A.C. Singh, 16-Q, R.H. Coats, 120 Parkdale Ave, Ottawa, ON K1A 0T6 avi.singh@statcan.ca

Abstract

We consider the problem of specifying the design-based error covariance structure in small area modeling with survey data because of too many unknown design parameters. As a compromise, it is customary to treat the estimated covariance as known although undesirable because of its instability, i.e., relative variance of this estimate could be high, and as a result, it could cause serious underestimation of variance of estimates. To alleviate this problem, one can either model the estimated error covariance structure in addition to the main task of modeling the estimated small area estimates (SAEs), or smooth the estimated covariance by only specifying its mean function. We advocate smoothing over modeling the error covariance because of strong simplifying assumptions needed in modeling the error covariance structure of estimated covariance of SAEs, and because practitioners, in general, prefer to take the least assumptions in modeling path philospphy. In practice, smoothing may work quite well as supported by the recent work of Singh, You, and Mantel (2005) who used the well known property of constant deff (design effect) over suitable subgroups of estimates to smooth the error covariance in modeling direct estimates from the cross-sectional and longitudinal data obtained from the Canadian Labour Force Survey. In this application, however, cross-sectionally the small areas were treated as strata, i.e., cross-sectionally, the error covariance was diagonal. This is rather restrictive because small areas or domains need not be strata. Moreover, using optimal estimating functions (EFs) of Godambe and Thompson (1986) for estimating model parameters, it can be shown that even when small areas are strata, additional summary statistics (in particular, the estimated domain count if it is not constant) for each domain or area should be used besides the direct SAEs, and thus the error covariance becomes necessarily nondiagonal. Use of EFs was also advocated by Singh, Folsom, and Vaish (2002, 2003) to define approximate EF-based Gaussian likelihood (EFGL) for obtaining efficient SAEs under a HB (hierarchical Bayes) framework. In this paper, we propose the idea of using g-deff (generalized design effect), defined earlier by Rao and Scott (1981) in the context of categorical data analysis, to deal with the case of nondiagonal error covariance. Simulation results for SAEs based on a linear mixed model show that the EFGL method with the proposed smoothing works quite well in general, and provides improved coverage of confidence intervals in particular.

Key Words: Estimating Functions; Generalized Deffs; Ignorable and nonignorable designs; unstable estimated error covarianc

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 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 because it depends on too many unknown quantities (e.g., the first and second order inclusion probabilities for the nonsampled units) that are not estimable. 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 $\hat{V}_t$ as well in addition to $t_{y,d}$ which renders the SAE problem even more compounded as variance of $\hat{V}_t$ now involves unknown third and fourth order sample inclusion probabilities. Under certain simplifying assumptions about the design, and the superpopulation model, the problem can, however, be simplified, see e.g., the use of 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 lot simpler than that of the error covariance structure, i.e., modeling the variance of $\hat{V}_t$. If the goal is to find a stable or smoothed estimate $\hat{V}_t$ of V_t without incurring too much bias, then one can simply specify the mean function for $\hat{V}_t$ and leave the covariance function unspecified, and then estimate parameters of the mean function by analogy with ordinary least squares. The smoothed estimate $\hat{V}_t$ is treated as known or may be allowed to depend on mean parameters of the model for $t_{y,d}$. It is advocated in this paper that smoothing might provide a reasonable practical compromise between the two extremes of modeling or no modeling of $\hat{V}_t$.

In this paper, we propose the use of 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_i '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 simple random sampling design or that the design is ignorable for the superpopulation model under consideration. In the case of diagonal $\hat{V}_t$ (i.e., when the small areas can be treated as strata), Singh, You and Mantel (2005) used the well-known property of constant deff ($V_{t(d)} / V_{t(d)}^*$) over suitable subgroups of direct estimates to smooth $\hat{V}_t$ by assuming that $\hat{V}_t^*$ is stable. More specifically, in the

context of the Canadian Labour Force Survey with partial overlapping samples over six months for each panel, they considered the combined longitudinal and cross-sectional data set of direct small area estimates for which the matrix $\hat{V}_t$ now becomes block diagonal where each block is a 6x6 matrix corresponding to each area. Now under the weak and rather plausible assumption of constant deff over six months for each area, and common fixed-lag correlations over areas and time, they got improved results over the alternative smoothing methods described in You, Rao, and Gambino (1999). However, the above method has some limitations: first, the nonnegativity of the smoothing matrix is not ensured when fixed lag correlations are replaced by averages over time and areas; second, for the more general case of areas not necessarily being strata as in the case of demographic domains, the matrix $\hat{V}_t$ becomes non-block-diagonal and it is not clear how the simple assumption of constant correlations can be applied to smooth the off-diagonal terms. In all these cases, however, the proposed method of g-deff can be used wherein the eigen-values are averaged over suitable subgroups to reduce their instability.

For the problem of SAE modeling with direct estimates $\{t_{y,d} : d = 1,...,D\}$, one can think of an implicit underlying superpopulation model at an aggregate level for the outcome variable y for each unit k in domain d; here the aggregate level model signifies that all the model covariates are at the area or domain (d) level. Conditional on the covariates, the model under the joint randomization of design (π) and superpopulation model (ξ, this is conditional on any random effects in the model) can be expressed for $d = 1,...,D$ as

$$t_{y,d} = T_{y,d} + e_{y,d},\ T_{y,d} = N_d A_{y,d}$$
$$A_{y,d} = A'_{x,d}\beta + \eta_d + N_d^{-1}\sum_k \varepsilon_{dk} \approx A'_{x,d}\beta + \eta_d = \mu_{y,d} \qquad (1.1)$$

$$e_{y,d} \sim N(0, V_{t(d)});\quad \eta_d \sim N(0, \sigma_\eta^2)$$
$$\varepsilon_{dk} \sim (0, \sigma_\varepsilon^2),\ N_d^{-1}\sum_k \varepsilon_{dk} \approx 0 \qquad (1.2)$$

where $T_{y,d}$ is the finite population total of y for area d; N_d is the known total count for area d which we will also denote as $T_{c,d}$, the total for the count variable c taking value 1 or 0 depending on whether the unit is in domain d or not; N is of course the whole population count; $A_{x,d}$ is the vector of domain-d averages for the p-vector of auxiliary variables x; β is a p-vector of fixed parameters, and η_d is the random effect corresponding to domain d assumed to be normal with mean 0 and variance σ_η^2; $e_{y,d}$ is the sampling or observation error in the estimated domain total and assumed to be normal with mean 0 and variance $V_{t(d)}$, and independent of η_d but may or may not be independent over d, and ε_{dk} is the superpopulation model error whose contribution to the small area total or average is negligible compared to other terms because N_d is assumed to be large. The model error ε_{dk} has mean 0 and variance σ_ε^2 but is not necessarily assumed to be normal although $t_{y,d}$ is assumed to be approximately normal due to central limit theorem provided n_d is sufficiently large. Note that $\mu_{y,d}$ is the superpopulation mean parameter of y given $A_{x,d}$ and is the limit of the corresponding finite population parameter $A_{y,d}$ as $N_d \to \infty$.

Now, the problem of smoothing $\hat{V}_t$ gets more complicated if one considers the above superpopulation model explicitly and the issue of optimal sample summary statistics to be used for estimating model parameters. In fact, the usual choice of summary statistics $\{ t_{y,d} : d = 1,..., D \}$ is somewhat ad hoc. Using Godambe and Thompson (1986), it can be shown (see next section) using census estimating functions that for aggregate level modeling, the optimal set of summary statistics is $\{ \psi_d = (t_{y,d}, t_{c,d}) : d = 1,..., D \}$. Here it is assumed that $t_{c,d}$ is not constant (i.e., not equal to N_d) as is typically the case in practice, else the usual set of summary statistics $\{ t_{y,d} : d = 1,..., D \}$ becomes optimal. Thus the error covariance $\hat{V}_{\psi(d)}$ for each d becomes a 2x2 matrix and the resulting overall matrix $\hat{V}_\psi$ necessarily nondiagonal requiring a more general smoothing method such as the proposed one.

Further, in the case of unit-level superpopulation model, the problem of smoothing $\hat{V}_\psi$ becomess considerably more involved because of additional summary statistics that enter in the picture from the theory of EFs. Note that Singh, Folsom, and Vaish (2002, 2003) proposed 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 in the interest of efficiency gains obtained from using more detailed information. 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 in the context of SAE modeling and the corresponding sample summary statistics. 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 the application of usual deff in the 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-smothed 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 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 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. Alternatively, Pfeffermann and Sverchkov (1999, 2003) proposed an elegant use of weighted likelihood approach for survey data. However, they require modeling of sampling weights which may not be desirable for practitioners who prefer the route of least assumptions in modeling as mentioned earlier. Using the idea of EFs, Godambe and Thompson (1986) showed how optimal EFs can be obtained under joint $\pi\xi$ - randomization. Thus for the aggregate level superpopulation model conditional on covariates $A_{x,d}$ and random effects η_d

$$y_{dk} = A'_{x,d}\beta + \eta_d + \varepsilon_{dk} \tag{2.1}$$

the optimal EFs for (η, β) are

$$\begin{aligned} \phi_{\eta(d)} &= \sum_k (y_{dk} - A'_{x,d}\beta - \eta_d) w_{dk} = t_{y,d} - (A'_{x,d}\beta + \eta_d) t_{c,d} \\ \phi_\beta &= \sum_d \sum_k A_{x,d} (y_{dk} - A'_{x,d}\beta - \eta_d) w_{dk} = \sum_d A_{x,d} t_{y,d} - \sum_d A_{x,d} (A'_{x,d}\beta + \eta_d) t_{c,d} \end{aligned} \tag{2.2}$$

where w_{dk} are the sampling weights, and for each domain d, k varies from 1 to n_d. It follows that the summary statistics $\{\psi_d = (t_{y,d}, t_{c,d})' : d = 1,...,D\}$ where $t_{y,d} = \sum_k^{n_d} y_{dk} w_{dk}$, $t_{c,d} = \sum_k^{n_d} w_{dk} = \hat{N}_d$, corresponding to the finite population quantities $\{(T_{y,d}, T_{c,d}) : d = 1,...,D\}$ form an optimal set for estimating the parameters (η, β). This is an interesting finding which came about by working with EFs under the superpopulation model. In contrast, with the traditional area-level model, one starts directly with $t_{y,d}$ as an estimate of the parameter of interest $T_{y,d} = N_d A_{y,d} \approx N_d (A'_{x,d}\beta + \eta_d)$, in which case it is easy to overlook the utility of the other summary statistic $t_{c,d}$ because the domain-level covariates N_d and $A_{x,d}$ are assumed to be known.

The aggregate-level model (1.1) can be generalized for the enlarged summary statistics ψ_d by introducing unknown incidental parameters p_d defined as the limit of N_d / N as $N \to \infty$. The model with all the parameters is given by

$$\begin{aligned} t_{y,d} &\approx N p_d \mu_{y,d} + e_{y,d} \\ t_{c,d} &\approx N p_d \mu_{c,d} + e_{c,d} = N p_d + e_{c,d} \end{aligned} \tag{2.3}$$

where $\mu_{y,d} = A'_{x,d}\beta + \eta_d$, $\mu_{c,d} = 1$. The sampling errors for $t_{y,d}$ in equations (1.1) and (2.3) do not change because $N p_d \approx N_d$. Observe that the introduction of incidental parameters p_d (although not of direct interest) is bit of an artifact but has the desired effect of engaging $t_{c,d}$ in estimating the parameters (β, η_d) as implied by EFs. However, unlike EFs, in the model (2.3), the data is separable from the parameters in that the mean function on the right hand side consists of only nonrandom covariates and unknown parameters. As a result, it becomes easier to apply the proposed method of g-deff for smoothing $\hat{V}_\psi$ unlike $\hat{V}_\phi$ corresponding to (2.2). Note that $\hat{V}_\phi$ involves the mean parameters and so smoothing would be required at each

cycle of MCMC in using EFGL. With model (2.3), the smoothing of $\hat{V}_\psi$ is done only once before starting the cycles of MCMC. However, unlike EFGL with ϕ, here the model's mean function is nonlinear in the enlarged parameter set $\{\beta, \eta_d, p_d, d = 1,..., D\}$ which would be computationally more demanding than the linear case. However, in the MCMC cycles for finding conditional posteriors, model mean turns out to be linear in (β, η_d) given p_d and so no new calculations are required except that there would be more steps in each MCMC cycle because of incidental parameters.

Next for unit-level modeling, (2.3) can be modified as follows:

$$\begin{aligned} \phi_{\eta(d)} &= \sum_k (y_{dk} - x'_{dk}\beta - \eta_d) w_{dk} = t_{y,d} - t'_{x,d}\beta - t_{c,d}\eta_d \\ \phi_\beta &= \sum_d \sum_k x_{dk} (y_{dk} - x'_{dk}\beta - \eta_d) w_{dk} = t_{xy} - t'_{xx'}\beta + \sum_d t_{x,d}\eta_d \end{aligned} \tag{2.4}$$

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 new set of summary statistics can be expressed with the new set of incidental parameters $\{p_d, \mu_{x,d}, \mu_{xx'}\}$ as :

$$\begin{aligned} t_{y,d} &\approx N p_d \mu_{y,d} + e_{y,d} \\ t_{c,d} &\approx N p_d + e_{c,d} \\ t_{x,d} &\approx N p_d \mu_{x,d} + e_{x,d} \\ t_{xy} &\approx N(\mu_{xx'}\beta + \sum_d p_d \mu_{x,d} \eta_d) + e_{xy,d} \\ t_{xx'} &= N \mu_{xx'} + e_{xx'} \end{aligned} \tag{2.5}$$

where, as defined earlier, $\mu_{y,d}$ is the limit of $A_{y,d}$ as $N_d \to \infty$ and $\mu_{x,d}$ is similarly defined. As in the case of aggregate-level, although the model (2.5) is now nonlinear because of incidental parameters, model mean turns out to be linear for computing conditional posteriors, and so no new computational complexity is involved except that there are more steps in 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 for domain counts or proportions for each covariate category.

Often, in practice, one may be interested in nonlinear modeling. Although it seems difficult to define optimal summary statistics for general nonlinear models at the unit level, it is indeed possible with categorical $x's$. Suppose, l denotes the l^{th} cell obtained by completely cross-classifying all the x-categories, and N_{dl} denotes the known population count for dl-subdomain. It can be shown from EFs that the optimal summary statistics ψ now are $\{t_{y,d}, t_{c,dl}, t_{xy}\}$ but there are too many $t_{c,dl}$ or $\hat{N}_{dl}$ because the total number of cross-classified l-categories can be very large. While it may be tempting to use all these summary statistics by introducing a large number of corresponding incidental parameters p_{dl}, it is best to avoid modeling with very unstable summary statistics. A compromise may be to collapse $t_{c,dl}$ over l to reduce the set of summary statistics. In doing so, we will only need to introduce unknown incidental parameters p_d as before, but we will also need to assume that the values of p_{dl} relative to p_d are known. This is a rather innocuous assumption because $p_{l|d}\ (= p_{dl}/p_d)$ can

be well approximated by the corresponding known finite population quantity N_{dl}/N_d. Now that the modeling problem is formulated in terms of the summary statistics ψ, we can turn to the question of smoothing $\hat{V}_\psi$ before model fitting using the EFGL method mentioned earlier.

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 $D \times 1$ 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 $\hat{V}_{\psi(d)} = \hat{\gamma}\hat{V}^*_{\psi(d)}$. 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

$$vec(\hat{V}_\psi) = \sum_j \lambda_j \, vec(\sum_k q_k q_k') + vec(e_v) \qquad (3.2)$$

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 statitics, 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),

$$V_\psi = Q\Lambda Q' = \sum_k^K \lambda_k q_k q_k' \qquad V_\psi^* = QQ' = \sum_k^K q_k q_k' \tag{3.3}$$

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

$$\hat{V}_\psi = \sum_j \hat{\lambda}_j \left(\sum_k q_k q_k'\right) \tag{3.4}$$

In the actual application of g-deff for smoothing, it may be better to avoid the effect of possibly different scaling of direct estimates on eigen-values. To do this, smooth parts of $\hat{V}_\psi$ expressed as a sandwich given by

$$\hat{V}_\psi = diag(\sqrt{\hat{V}_{\psi(k)}})\, R(\hat{\rho})\, diag(\sqrt{\hat{V}_{\psi(k)}}) \tag{3.5}$$

where $R(\hat{\rho})$ is the correlation matrix in the middle, and $\hat{\rho}$ denotes the estimated vector of $K(K-1)/2$ correlations. Now we can use the deff-based GVF (equation 3.1) to smooth the left and right parts of the sandwich, while for the middle part, the g-deff based GVF (equation 3.2) can be used to extract the smoothed correlation vector $\hat{\rho}$ from the corresponding smoothed matrix $\hat{R}$ (this need not be a correlation matrix), and then obtain $\hat{V}_\psi$ as

$$\hat{V}_\psi = diag(\sqrt{\hat{V}_{\psi(k)}})\, R(\hat{\rho})\, diag(\sqrt{\hat{V}_{\psi(k)}}) \tag{3.6}$$

Some additional benefits of using the above sandwich form for smoothing are that it allows for more freedom in choosing parameters for smoothing because we have K deffs and K eigen-values based on the matrix $R(\hat{\rho})$ as opposed to only K eigen-values or g-deffs based on the matrix $\hat{V}_\psi$, and that the two outside parts of the sandwich can be allowed to depend on model mean parameters as mentioned earlier under section 3.1. Note that in the g-deff based smoothing (3.4), it is not clear how one can make the eigen-value or the eigen-vector depend on unknown mean parameters.

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 modeling, $\hat{V}_\psi$ 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

$$\hat{V}_{\psi} = \begin{pmatrix} A & B \\ B' & C \end{pmatrix} \tag{3.7}$$

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

$$\begin{pmatrix} A & B \\ B' & C \end{pmatrix} = H^{-1} \begin{pmatrix} A & O \\ O & C - B'A^{-1}B \end{pmatrix} H^{-1\prime}, \quad H = \begin{pmatrix} I & O \\ -B'A^{-1} & I \end{pmatrix} \tag{3.8}$$

where $C - B'A^{-1}B$ 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 $\hat{A}$ (using the g-deff idea) which being block-diagonal is not computationally demanding, and then obtain $\hat{V}_{\psi}$ 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 $d = 1, \ell\ , D$ 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 $k = 1, \ell\ , N_d$. The covariate $x_{dk} = \upsilon_d + \delta_{dk}$ where $\upsilon_d \sim N(0, 0.1)$ and $\delta_{dk} \sim N(0, 1)$. We generate $M = 150$ population level data sets with common x_{dk} and N_d where N_d's are generated using N_0 =3000. Note that the substratum sizes vary over the 150 populations. 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 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 150 populations, we generate the corresponding 150 samples. In our simulation experiment, $(n_{d+}, n_{d-}) = (2,10), (5,15)$, 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. Also, we didn't consider the sandwich

-type g-deff method (3.5) for smoothing $\hat{V}_\psi$ but used the initial g-deff method (3.4). To avoid computational complexity, we used the smoothing based on the decomposition (3.8) 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 as (estimation of the parameter σ_ε^2 was not considered as it was not needed)

$$\beta \sim U(R), \quad \sigma_\eta^2 \sim IG\left(\nu_0/2, \sigma_{\eta_0}^2/2\right) \tag{4.1}$$

where $\sigma_{\eta_0}^2 = .01$, a small positive number, and $\nu_0 = 3$, the smallest possible integral value greater than 2 which gives a practically meaningful interpretation of the posterior mean of $[\sigma_\eta^2 | \eta]$ obtained as a weighted average

$$E\left[\sigma_\eta^2 \mid \eta\right] = \left(D\left(\sum_d \eta_d^2/D\right) + (\nu_0 - 2)\sigma_{\eta_0}^2\right) \Big/ (D + \nu_0 - 2) \tag{4.2}$$

The results from the simulation study are presented in Tables 1 and 2. In terms of point estimation and Standard deviation, 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 and length of prediction intervals, the unsmoothed method EFGL-u starts showing deterioration as seen from Table 2. It is expected that variance of SAEs would suffer from downward bias due to underestimation of their variances by regarding $\hat{V}_\psi$ or its smoothed version $\hat{V}_\psi$ as known. However, it is interesting to note that for smaller sample sizes, there is marked difference between the coverage probability and interval width between the smoothed and unsmoothed methods and smoothing is seen to correct in the right direction. Nevertheless, when sample sizes for small areas are rather low as seen in the lower half of Table 2, even after smoothing, there is significant undercoverage of prediction intervals. This suggests that $\hat{V}_\psi$ is still not sufficiently smooth and perhaps collapsing of small areas with similar random effects as suggested in Singh, Folsom and Vaish (2003) might alleviate this problem. In future, this and other issues related to the performance of smoothing methods for other choices of area sample sizes will be investigated further.

5. Summary and Concluding Remarks

In this paper we considered the problem of smoothing the error covariance 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 error covariance, and 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 correct underestimation 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.

Finally, it is interesting to note that the use of GVF was also identified by Eltinge et. al (2002) in trying to stabilize the covariance matrix in the context of quadratic-score type tests for model checks in analysing survey data, and considered their relationships with Rao-Scott corrections to chi-square tests. However, they didn't consider the use of g-deff in GVF-type modeling that seems to provide a simple but important approach for the problem of SAE considered in this paper.

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.

Eltinge, J.L., Cho, M.J., and Hinrichs, P. (2002). Use of Generalized variance functions in Multivariate Analysis, *Proceedings of American Statistical association, Section on Survey Research Methods.*

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. (1960). A new approach to sampling from finite populations I, II (with discussion), *JRSS (B)*, 28, 310-328.

Godambe and Thompson, M.E, (1986), "Parameters of Super population and Survey Population, Their Relationship and Estimation," *International Statistical Review,* **54**, pp. 127-38.

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. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya: The Indian Journal of Statistics*, Ser. B, **61**, 166-186.

Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In: *Analysis of Survey Data*, eds. C.J. Skinner, and R.L. Chambers, New York: Wiley, 175-195.

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 honor 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 Statistiscal Policy Working paper* #36, pp. 47-74 (www.fcsm.gov/working-papers/)

Singh, A.C., You, Y., and Mantel, H.J. (2005). Use of generalized design effect for variance function modeling in small area estimation from survey data, *Statistical Society of Canada, 33rd Annual Meeting,* Saskatoon, Saskatchewan, June 12-15.

You, Y, Rao, J.N.K., and Gambino, J.G. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach, *Survey Methodology,* 29, 25-32.

Table 1: Average Posterior Mean and Standard Deviation for Model Parameters: Nonignorable Sample Design

$(n_{d+}, n_{d-}) = (5,15)$

Parameter (True Value)	Average Posterior Mean		Average Posterior Standard Deviation	
	EFGL-u	EFGL-s	EFGL-u	EFGL-s
$\beta_0(0.5)$	0.4927	0.4933	0.0478	0.0477
$\beta_1(1.0)$	0.9984	0.9983	0.1225	0.1226
$\sigma_\eta^2(0.2)$	0.1970	0.1963	0.0320	0.0320

$(n_{d+}, n_{d-}) = (2,10)$

Parameter (True Value)	Average Posterior Mean		Average Posterior Standard Deviation	
	EFGL-u	EFGL-s	EFGL-u	EFGL-s
$\beta_0(0.5)$	0.4816	0.4790	0.0505	0.0504
$\beta_1(1.0)$	0.9852	0.9839	0.1011	0.1030
$\sigma_\eta^2(0.2)$	0.2048	0.1980	0.0347	0.0345

Table 2: 95% Coverage Probability and Ratio of Prediction Interval (PI) Widths: Nonignorable Sample Design

$(n_{d+}, n_{d-}) = (5,15)$

Percentiles and Means over Small Areas	Coverage Probability		Ratio of Average PI Widths
	EFGL-u	EFGL-s	EFGL-s/EFGL-u
75%	0.9333	0.9533	1.02
Mean	0.9183	0.9386	1.01
25%	0.9000	0.9267	1.00

$(n_{d+}, n_{d-}) = (2,10)$

Percentiles and Means over Small Areas	Coverage Probability		Ratio of Average PI Widths
	EFGL-u	EFGL-s	EFGL-s/EFGL-u
75%	0.8433	0.8833	1.06
Mean	0.8252	0.8659	1.06
25%	0.8067	0.8467	1.04