Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 16 No. 1, 2018 (New Series), pp 399-409

Variance Modeling for Domains

Sam Hawala[1] **and Partha Lahiri**[2]
[1]*Resonate, Reston, VA, USA*
[2]*University of Maryland, College Park, MD, USA*

Received: April 22, 2018; Reviewed: September 10, 2018; Accepted: September 15, 2018

Abstract

In order to reduce both computational and publication burdens, Generalized Variance Function (GVF) was introduced in official statistics production. Subsequently, the GVF found its way into the production of smoothed variance estimates of survey-weighted estimates for small areas, which are inputs in area level models. Parallel to the GVFs, a different approach, which we call variance modeling, was introduced in small area estimation. While the concept was introduced in an integrated area level model in order to account for an additional level of uncertainty in making inferences on small area means, such a model has the ability to make inferences about the small area sampling variances as well. In this paper, we develop a theoretical framework for variance modeling of survey-weighted count estimates for domains with the goal to estimate design-based variances of survey weighted count estimates for domains — small and large. We then propose a few variance models and compare them using the American Community Survey (ACS) data.

Key words: Area level modeling; design-based; design-effect; effective sample size; generalized variance function (GVF); small area estimation.

1 Introduction

National statistical offices routinely publish numerous tables of estimates on different socio-economic, health, and other measures to keep the public informed about the well-being of a nation. There is also a great demand to report certain estimated measures of uncertainty such as variance estimates so that reliability of the published estimates can be ascertained. Computation of the variance estimates by a traditional survey method (e.g., Cochran 1977) generally imposes heavy computational as well as publication burden on the national statistical offices. Generalized Variance Function (GVF) method has been proposed to address this issue. Variables are placed into groups such that they are *similar* within a group in terms of intra-class correlation or design effects. A GVF is then proposed for each group by fitting a model that relates design-based variance estimates to the corresponding survey-weighted estimates and possibly some design variables. The GVFs then replace numerous variance estimates to save publication costs of outputs from the

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Corresponding author: Partha Lahiri
Email: plahiri@umd.edu

survey used to develop GVFs and both publication and computation costs of outputs for future surveys. Researchers have also found the use of GVF in improving traditional variance estimators for domains; see, e.g., Eltinge et al. (2000), Elinge and Sukasih (2001) and Hinrichs (2003). For ongoing surveys, GVFs can be updated from time to time. For a good review of the GVF method, the readers are referred to Wolter (1985, Chapter 5). Researchers at the U.S. Census Bureau have been using GVF for the Current Population Survey (CPS) since 1947 (see, Hansen et al. 1953).

The use of GVF in small area estimation can be traced back to the pioneering paper by Fay and Herriot (1979), who proposed an empirical Bayes method for producing per-capita income estimates for small areas. Using standard survey-weighted estimates and their associated design-based variance estimates for a few large areas, they developed a GVF that establishes a relationship between the design-based variance estimate of survey-weighted per-capita income for an area and the corresponding population size estimate and the survey-weighted per-capita estimate. They assumed the GVF thus developed for large areas to hold for small areas and then used this synthetic assumption to motivate an approximate relationship between the true sampling variance of survey-weighted per-capita income estimator and the true per-capita income for small areas. They used such an empirical relationship in suggesting a variance stabilizing transformation of the survey-weighted estimator of per-capita income for their sampling model, which simplified their resulting empirical Bayesian methodology considerably. We note that a reasonable variance stabilizing transformation can be also motivated without using GVF; see, for example, Carter and Rolph (1974), Efron and Morris (1975), Jiang et al. (2001), Raghunathan et al. (2007), and Casas-Cordero et al. (2016).

For small areas (domains), traditional design-based variance estimation techniques are known to be unreliable due to small sample sizes in the areas; see Rao and Molina (2015) for a comprehensive review of different small area estimation theory and applications. While GVF was initiated to address the computational and publication issues, its potential utility to improve on the traditional variance estimates for small areas cannot be overemphasized. However, there are a number of challenges in developing reasonable GVF for small areas because of noisy survey-weighted estimates and their associated design-based variance estimates. We note that Fay and Herriot (1979) and others who used variance stabilizing transformations on survey-weighted estimators (e.g., Casas-Cordero et al. 2016) did not, however, explicitly discuss the issue of estimation of design-based variance of a survey-weighted estimator. A suitable method applied on their hierarchical models may lead to variance estimators that are more reliable than the traditional design-based variance estimators for some small area estimation problems, but this claim needs further research.

Maples et al. (2009) and others extended the GVF approach for estimating sampling variance of a survey-weighted estimator for small areas. Unlike Fay and Herriot (1979), the method proposed by Maples et al. (2009) does not restrict their GVF fitting based on a subset of large areas and incorporates sampling errors of the traditional design-based variance estimator and borrows strength from relevant area specific auxiliary variables. To elaborate, they examined the sampling distribution of the design-based variance estimators for the logarithm of survey-weighted counts. Maples et al. (2009) suggested a parametric model applied directly to the design-based direct estimates and claimed reasonable performance in their Small Area Income and Poverty (SAIPE) application. Earlier Otto and Bell (1995) considered a multivariate version where sampling co-

variance matrices were also directly modeled using Current Population Survey (CPS) data. In the context of consumer expenditure survey, Hinrichs (2003) proposed both univariate and multivariate hierarchical models to improve the GVF methodology for variance estimation for domains.

In the context of a small area estimation problem associated with the U.S. consumer expenditure survey, Arora and Lahiri (1997) proposed a joint modeling of means and variances in order to capture different sources of variations. Though not discussed in their paper, Bayesian estimates of sampling variances of direct survey-weighted estimators can be produced in a straightforward way using the Monte Carlo Markov Chain (MCMC) applied on their joint hierarchical model. We stress that this approach is different from the traditional GVF or its extensions described in the previous paragraph as design-based variances are not directly modeled. To elaborate, Arora and Lahiri (1997) considered a multiplicative model for sampling variance of the direct survey-weighted estimator in which the first factor is a known function of different area specific design variables and the second factor is an area specific random variance component on which a hierarchical model is assumed. Similar ideas can be found in You (2002). In some cases, area specific random variance components can be modeled as a function of means, which simplifies the modeling. For example, in the context of estimating small area proportions, different functions of unknown area specific random proportion have been proposed for the area specific random variance component; see, e.g., Liu et al. (2007, 2014), Hawala and Lahiri (2010), Maples (2016).

In this paper, we propose a simple method to estimate the sampling variances of survey-weighted counts for domains – small or large. This calls for first modeling the sampling variance of survey-weighted count for a domain as a known function of domain specific unobserved true proportion, domain specific observed design variables and unknown global parameters. The global parameters can be estimated by fitting the assumed model for similar large domains. One may substitute the unobserved true proportion in the assumed variance model by its survey-weighted estimate to obtain variance estimate for large domains. For small areas, a hierarchical model for the unobserved area specific proportions can be assumed in order to borrow strength from different area specific auxiliary data sources and to explain different sources of variations. Under a Bayesian approach, MCMC can be applied on this hierarchical model to produce the posterior distribution of true unknown sampling variances, which can be used for all inferential purposes for the true unknown sampling variancse, including point estimation.

In section 2, we introduce notations used in the paper. In section 3, we first provide an analytical argument for inclusion of certain design factors in variance modeling for a sample design in common use. In this section, we also discuss different possible models for sampling variances of survey-weight counts and the estimation of global parameters. In section 4, we compare different variance models using data from the American Community Survey. We end the paper by offering a few concluding remarks and prospect for future work in this area.

2 A Table of Notations

The following notations are for a given domain i. For notational simplicity, we supress the domain index i for developing a variance model for the domain.

U: the set of all units (e.g., persons) in the survey population for the domain,

N: number of units in U,

For unit $k \in U$,

$$y_k = \begin{cases} 1 & \text{if unit } k \text{ satisfies an attribute of interest (e.g., a binary poverty status),} \\ 0 & \text{otherwise,} \end{cases}$$

$Y = \sum_{k \in U} y_k$, total number of units in U satisfying the attribute, the main parameter of interest,

$P = \frac{Y}{N}$, proportion of units in U satisfying the attribute,

s: a sample of all units,

s_h: set of all units nested within a larger compound unit h in s (e.g., the compound unit could be a household if the unit is a person),

N_h: number of units in s_h,

$n = \sum_h N_h$: number of units in s, where $\sum_h$ denotes summation over all compound units in s,

w_k: survey weight associated with unit k in s,

$\hat{Y} = \sum_{k \in s} w_k y_k$, survey-weighted estimator of Y,

$V \equiv V(\hat{Y})$: true design-based variance that incorporates all sources of variabilities, including variability due to design and weighting,

$\hat{V} \equiv \hat{V}(\hat{Y})$: a suitable estimator of $V(\hat{Y})$ (e.g., Fay's successive difference replication variance estimator; see Fay and Train, 1995),

$\hat{N} = \sum_{k \in s} w_k$,

$\hat{P} = \hat{Y}/\hat{N}$.

We use lower case letters to indicate the value of a given estimator from a given sample. For example, we obtain $\hat{p}$ from $\hat{P}$ for a given sample.

3 Variance Modeling

In this section, we first illustrate modeling of a binary variable for a simple design. We then discuss how such a model helps proposing different variance models that can be used to produce variance estimates for domain survey-weighted estimator $\hat{Y}$. To this end, we introduce a binary variable z_h for the compound unit $h \in s$. We assume the following model:

Model M: $y_k = z_h, \forall k \in s_h$ with $E(z_h) = \pi_h, V(z_h) = \sigma_h^2 = \pi_h(1 - \pi_h)$ and

$$Cov(z_h, z_{h'}) = 0, \forall h \neq h', \quad (3.1)$$

where π_h can be interpreted as a superpopulation true proportion of units satisfying the attribute in the compound unit h. This assumption that z_h's are uncorrelated across h may be reasonable for a design with mild effect due to possible clustering of the compound units (e.g., design used in the American Community Survey). The assumption $y_k = z_h, \forall k \in s_h$ can be justified for some variables. For example, if y_k denotes the poverty status of person k in household h, then the assumption is justified if poverty status of a person is determined by that of the household where he/she lives. Such an example can be found in National Socioeconomic Characterization Survey (CASEN) CASEN in Chile (see, e.g., Casas-Cordero et al. 2016), American Community Survey in USA, and surveys in other countries.

Noting that $\hat{Y} = \sum_h z_h(\sum_{k\in s_h} w_k)$, we obtain $V_M(\hat{Y})$, variance of $\hat{Y}$ under model M, as

$$V_M(\hat{Y}) = n\left[\sum_h q_h\pi_h(1-\pi_h)\right]\sum_h \gamma_h \frac{(\sum_{k\in s_h} w_k)^2}{N_h}, \tag{3.2}$$

where

$$q_h = \frac{N_h}{\sum_h N_h}, \quad \gamma_h = \frac{N_h\pi_h(1-\pi_h)}{\sum_h N_h\pi_h(1-\pi_h)}. \tag{3.3}$$

Using the concavity of the function $f(x) = x(1-x),\ 0 \le x \le 1$, and the Jensen's inequality, we obtain the following upper bound from equation 3.2

$$V_M(\hat{Y}) \le n\pi(1-\pi)\sum_h \gamma_h \frac{(\sum_{k\in s_h} w_k)^2}{N_h}, \tag{3.4}$$

where $\pi = \sum_h q_h\pi_h$. We can interpret π as a superpopulation proportion of units in the domain satisfying the attribute. Let $n_* = n \times$ deff, where deff $= \sum_h(\sum_{k\in s_h} w_k)^2/n$ denotes the design effect. The upper bound in equation 3.4 reduces to $n_*P(1-P)$ when γ_h and π are substituted by q_h and P, respectively. This motivates the following variance model:

$$V_1 \equiv V_1(P; n_*) = n_*P(1-P). \tag{3.5}$$

Note that the model M does not incorporate the possible variability due to the weighting process. The upper limit may take care of this additional variation, but it may be subject to bias. In order to reduce bias, one may apply a multiplicative factor b to 3.5 and obtain the following alternative variance model:

$$V_2 \equiv V_2(P; n_*) = bn_*P(1-P) = dP(1-P), \tag{3.6}$$

where $d = b \times n_*$ and b is the average size of the compound unit. It may be possible to correct for the bias by considering a more complex variance model that involves P and different design factors such as $n,$ deff, and one or more calibration factors b. We shall now elaborate on this.

Bringing back the domain index i, the variance model 3.5 motivates the following variance model for domain i:

$$V_{1i} \equiv V_{1i}(P_i; d_i) = n_{i;*} P_i (1 - P_i), \tag{3.7}$$

where $n_{i;*} = n_i \times \text{deff}_i$ and $\text{deff}_i = \sum_h (\sum_{k \in s_{ih}} w_k)^2 / n_i$, the summation being over all sampled compound units in domain i. The design effect deff_i could be free of i, especially for small area (domain) applications.

We propose Model I-III and Model V to correct for possible bias in the variance model 3.5. Model IV is used by the U.S. Census Bureau.

$$\begin{aligned}
\textbf{Model I}: \quad V_{i1} &= \exp(b_0) P_i^{b_1} (1 - P_i)^{b_2} (d_i)^{b_3}, && (3.8)\\
\textbf{Model II}: \quad V_{i2} &= \exp(b_0) P_i^{b_1} (1 - P_i)^{b_2} a_i^{b_3}, && (3.9)\\
\textbf{Model III}: \quad V_{i3} &= \exp(b_0) [P_i(1 - P_i)]^{b_1} (d_i)^{b_3}, && (3.10)\\
\textbf{Model IV}: \quad V_{i4} &= b_1 Y_i + b_2 Y_i^2, && (3.11)\\
\textbf{Model V}: \quad V_{i5} &= d_i P_i (1 - P_i), && (3.12)
\end{aligned}$$

where $d_i = b n_{i;*}$.

Remark 1: Note that in the above the coefficients ($b_0,\ b_1,\ b_2$ and b_3) have different interpretations across different models. We use the same notation for simplicity.

Remark 2: In a logarithmic scale, Model I-III and Model V can be viewed as a multiple linear regression model. In the log-transformed Model I, the independent variables are $\log P_i,\ \log(1 - P_i)$, and $\log d_i$. Model II, motivated from Maples et al. (2009), is obtained from Model I when $\log(d_i)$ is replaced by $\log a_i$, where a_i is the number of compound units in the sample for domain i. Note that $\log P_i$ and $\log(1 - P_i)$ are likely to be correlated. So in order to avoid the possible multicollinearity problem, we consider Model III where we replace the two independent variables of Model I, i.e., $\log P_i$ and $\log(1 - P_i)$ by a single independent variable $\log[P_i(1 - P_i)]$. In the log-transformed Model V, the coefficients of all the independent variables are 1.

Remark 3 To estimate the model parameters $b_0,\ b_1,\ b_2$, and b_3 for Model I-III, one may use standard design-based variance estimates $\hat{V}_i$ for the dependent V_i and survey-weighted estimates p_i for P_i for similar large domains and use least squares method for the models in the logarithmic scale. The fitted model can be used for producing variance estimates for similar items. Small area (domain) model fitting and variance estimation are challenging because the direct estimates of both V_i and P_i are unreliable due to small domain sample sizes. For small domains, one may estimate the model parameters using estimates for large domains. The sampling variance may then be viewed as a function of P_i and so Bayesian method using an appropriate hierarchical model such as the models used by Liu et al. (2007, 2014) can be used to make inferences about the sampling variance of survey-weighted estimator.

4 Evaluation of Different Variance Modeling using American Community Survey Data

For evaluation of our proposed methodology, we treat U.S. states. We choose these domains so standard survey-weighted estimates and their associated Fay's successive difference replication variance estimates are reliable. Model V is the simplest adjustment of 3.7. So it will be interesting to compare variance estimates resulting from Model V with the Fay's successive difference replication variance estimates. Figure 1 provides a scatter plot of these two estimates for the states. In the plot, we denote estimates from Model V by V_{approx} and Fay's estimates by V.

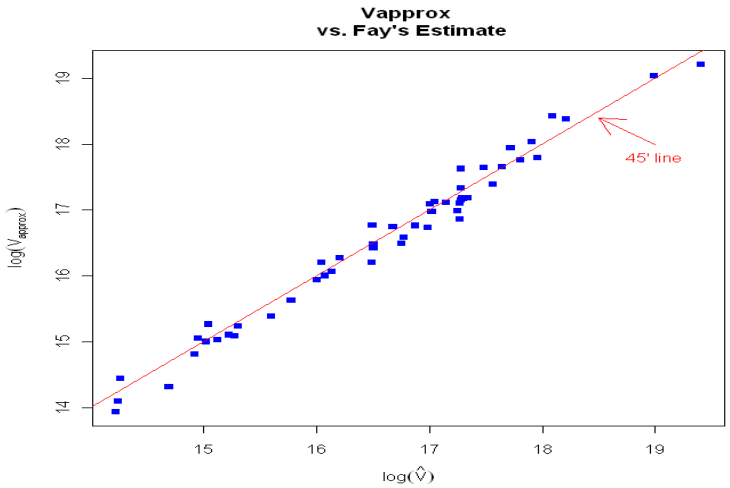

Figure 1: Scatter plot of V_{approx} obtained from Model V vs. the Fay's variance estimate for the U.S. states

From Figure 1, we observe that V_{approx} approximates the Fay's estimates well. This suggests that Model V is reasonable because at the state level the Fay's method usually provides reasonable variance estimates. Since Models I-III are all in the same logarithmic scale in the same dependent variable, we can compare these three models in terms of the usual model selection criteria. Table 1 displays these model selection statistics. For a review of model selection, the readers are referred to the IMS monograph edited by Lahiri (2001).

Table 1: Models on Log $(\hat{V})$:

Criteria	Model I	Model II	Model III
Adj. R^2	0.9735	0.9019	0.9741
AIC	-14.90	51.90	-16.88
BIC	-5.24	61.56	-9.15
PRESS	2.14	7.82	2.07
R_{PRESS}	0.9709	0.8939	0.9719

From Table 1, we can see that Model I and III both outperform Model II in terms of the model selection statistics considered. Model III appears to be slightly better than Model I. This suggests that the design factor d_i is a reasonable component of the variance model. We cannot use the model selection criteria given in Table 1 to compare variance estimates motivated from Model I-III with the variance estimates from Model IV since the dependent variables are in different scales.

To compare all the five models, we compute relative differences of the variance estimates from the Fay's estimate for all the 50 states and the District of Columbia. We define the relative difference as

$$RD = (\tilde{v} - \hat{v})/\hat{v},$$

where $\tilde{v}$ is one of the five variance estimates obtained from the five GVF models. Figure 2 displays the box-plots for each of the five models. Model I and Model III emerge as the two best performers. Model II and IV seem to have some underestimation problem. The variance estimate derived from Model V, i.e. $\bar{b}\hat{V}_{approx}$ (Figure 2) seems promising, although there is a tendency for possible overestimation. This conservative approach may be reasonable when we do county level estimation as we do not know how good the Fay's variance estimator is in terms of capturing all sources of variabilities from the sampling and weighting processes.

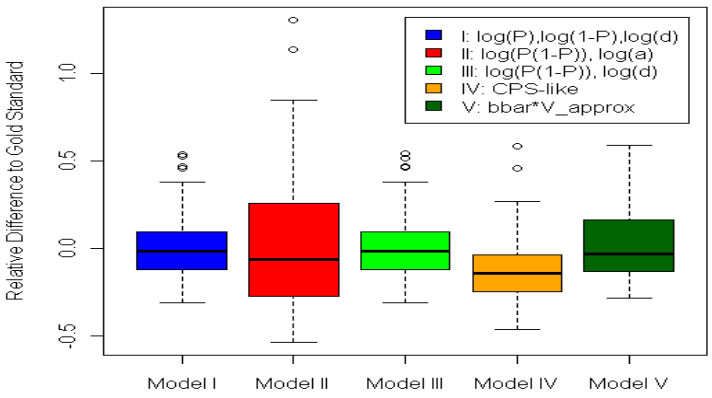

Figure 2: Comparisons of variance estimates based on Relative Difference

5 Concluding Remarks

In this paper, we propose different variance models for survey-weighted domain totals. The theoretical framework suggested in the paper could be potentially extended to deal with non-binary data and for more complex designs such as the design considered in Gabler et al. (1999). This could be a good future research topic. Small area variance modeling is a challenging research topic. The synthetic approach proposed in this paper could introduce considerable bias in the variance

estimates, if the synthetic assumptions are not reasonable. It may be possible to reduce the bias of the synthetic variance estimation bias by using an integrated model that combines a synthetic model such as the one proposed in this paper for large areas with a more elaborate hierarchical model such as the ones considered by Liu et al. (2007, 2014) for the small areas. We plan to pursue this research area in the future.

6 Acknowledgements

The research of the second author was supported in part by the National Science Foundation Grant Number SES-1534413. The computation was completed when the first author was an employee and the second author is a Fellow at the U.S. census Bureau.

References

Arora, V. and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica,* **7**, 1053-1063.

Carter G.M. and Rolph J.E. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, **69**, 880-885.

Casas-Cordero, C., Encina, J. and Lahiri, P. (2016). *Poverty Mapping for the Chilean Comunas, In Analysis of Poverty Data by Small Area Estimation*. ed. Monica Pratesi, Wiley Series in Survey Methodology.

Eltinge, J.L. and Sukasih, A. (2001). *Approximation methods for covariance matrix estimators using analysis of diary and interview data from the U.S. consumer expenditure survey*. BLS document.

Eltinge, J.L., Sukasih, A. and Weber, W. (2000). *Feasibility of constructing combined estimators using consumer expenditure interview and diary data*. Paper presented to the Bureau of Labor Statistics Conference on Issues in measuring Price Change and Consumption.

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of American Statistical Association*. **70**, 311-319.

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**, 269-277.

Fay, R.E., and Train, G. (1995). *Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties*. Proceedings of the Government Statistics Section, Alexandria, VA: American Statistical Association, pp 154-159.

Gabler, S., Haeder, S., and Lahiri, P. (1999). A model-based justification of Kish's formula for design effects for weighting and clustering, *Survey Methodology*, **25**, 105-106.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory.* 2 Volumes. New York: John Wiley and Sons.

Hawala, S. and Lahiri, P. (2010). Variance Modeling in the U.S. Small Area Income and Poverty Estimates Program for the American Community Survey, *Proceedings of the American Statistical Association, Survey Research Section,* 4655-4663.

Hinrichs, P. (2003), *Consumer Expenditure Estimation incorporating Generalized Variance Functions in Hierarchical Bayes Models*. Ph.D. Dissertation, Department of Mathematics and Statistics, University of Nebraska, Lincoln.

Jiang J, Lahiri P, Wan S, Wu C. (2001). *Jackknifying in the Fay-Herriot model with an example*. Proceedings of the Seminar on Funding Opportunity in Survey Research, Council of Professional Associations on Federal Statistics, Arlington, VA, 2001, 75-109.

Lahiri, P. (2001). *Model Selection*. Lecture Monograph, Volume 38, Lahiri, P. ed.

Liu, B. (2009). *Hierarchical Bayes estimation and empirical best prediction of small area proportions*. Ph.D. Dissertation, Joint Program in Survey Methodology, University of Maryland, College Park.

Liu, B., Lahiri, P., and Kalton, G. (2007). *Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions*. Proceedings of the American Statistical Association, Survey Research Section, 3181-3186.

Liu, B., Lahiri, P. and Kalton, G. (2014). Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions, *Survey Methodology*, **40**, 1-13.

Maples, J., Bell, W., and Huang, E.T., (2009). *Small Area Variance Modeling with Application to County Poverty Estimates from the American Community Survey*. Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association, 5056-5067.

Maples, J.J. (2016). *Estimating Design Effects in Small Areas/Domains through Aggregation*. Proceedings of the Section on Survey Research Methods, Alexandria, VA, American Statistical Association, 670-681.

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

Raghunathan, R.E., Xie, D., Schenker, N., Parsons, V.L., Davis, W.A., Dodd, K.W., and Feuer, E.J. (2007). Combining Information From Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening, *Journal of the American Statistical Association,* **102**, 474-486, DOI 10.1198/016214506000001293.

Singh, A.C., Folsom, R.E. JR., and Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. Federal Committee on Statistical Methods Statistical Working Paper, No. 39, (`http://www.fcsm.gov/05papers/Singh_etal_IIIC.pdf`).

Tomlin, P. (1974). U.S. Census Bureau memorandum "Justification of the Functional Form of the GATT Curve and Uniqueness of Parameters for the Numerator and Denominator of Proportions," from Paul Tomlin to the Record, dated December 11, 1974.

U.S. Census Bureau, (2006). Design and Methodology: American Community Survey, U.S. Government Printing Office, Washington, DC. `http://www.census.gov/acs/www/Downloads/tp67.pdf`*BOC 2006* .

Valliant, R. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling. *Journal of the American Statistical Association,* **82**, 499-508.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for subprovincial areas of Canada. *Survey Methodology*, **34**, 19-27

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.