Variance Modeling in the U.S. Small Area Income and Poverty Estimates Program for the American Community Survey *

Sam Hawala, U.S. Census Bureau
Partha Lahiri, University of Maryland, College Park

September 16, 2010

Abstract

In small area income and poverty estimates program of the U.S. Census Bureau (SAIPE), one of the challenges is the estimation of sampling variances of the direct survey weighted estimators for the counties. Design-based methods can be highly unreliable primarily because of small sample sizes in the area. Generalized variance function (GVF) methods have been previously used in the SAIPE and other small area estimation projects to obtain smoothed variance estimates. In the context of county level estimation of the number of school-age (5-17 year old) children in poverty using the American Community Survey (ACS) data, we propose a new approach in which a person level working model is used to motivate a GVF. The model fitting and model comparison are done at the state level where the design-based estimates and the corresponding standard direct design-based variance estimates are assumed to be reliable. The proposed GVF model is an important component of a larger multilevel model that can be used in the future to produce improved estimates of different parameters of interest.

1 Introduction

Survey statisticians have long been interested in modeling design-based variance of a survey estimator as a function of its design-based expectation. The researchers at the U.S. Census Bureau have been using such variance modeling for the Current Population Survey (CPS) since 1947 (see Hansen, Hurwitz, and Madow 1953). The main use of such variance model, referred to as the Generalized Variance Function (GVF) in the sample survey literature, has been in reducing the computational and publication burden in variance estimation from large scale sample surveys in which the users are interested in many different survey items for many subgroups of the survey population. The model is often assumed based on visual inspection of plots of design-based sampling variance estimates against the survey-weighted estimates for a few items or by gaining some insight from the design-based sampling variance formula. The model parameters are estimated for certain groups of items with *similar* intra-class correlation or design effects. For a good review of the GVF method, the readers are referred to Wolter (1985, Chapter 5).

The use of the GVF in small area estimation has a relatively shorter history. Fay and Herriot (1979) are probably the first to introduce such a method in a complex survey setting in order to motivate the sampling error component of their two-level Bayesian model. Their area level model, used to estimate the per-capita income for small places, can be described as follows:

*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.

The Fay-Herriot Model:

For area $i = 1, \cdots, m$,

Level 1: $\hat{\theta}_i | \theta_i \overset{ind}{\sim} [\theta_i,\ D_i]$,

Level 2: $\theta_i \overset{ind}{\sim} \left[\boldsymbol{x}_i^T \boldsymbol{\beta}, A\right]$,

where $\hat{\theta}_i = \log(y_i)$, $\theta_i = \log(\mu_i)$; y_i is the survey-weighted estimate of the true per-capita income μ_i; $D_i = 9/N_i$, where N_i is the known population size; $\boldsymbol{x}_i$ is a $p \times 1$ vector of known fixed auxiliary variables; $\boldsymbol{\beta}$ and A are unknown model parameters. In the above model, Level 1 was used to describe the sampling error distribution of the the log-transformed per-capita survey-weighted estimates. Level 2 was used to *borrow strength* by relating logarithm of the true per-capita income θ_i to various area level administrative and census data contained in $\boldsymbol{x}_i$.

The assumption of known sampling variances D_i was motivated using a GVF obtained empirically. Using data from eight states, Fay and Herriot (1979) first obtained an empirical relationship: $cv_i \approx 3/\sqrt{N_i}$, where cv_i is the estimated coefficient of variation of y_i with its variance estimated using a standard design-based variance estimation technique. They then made a synthetic assumption that the slope of the regression, i.e. 3, remains the same for all small areas, small or large, and concluded that the true sampling variance of y_i is given by $D_i \mu_i^2$, which suggested their log-transformation of y_i to stabilize the variance. Fay and Herriot (1979) obtained their empirical Bayes estimator of θ_i and then used a simple back-transformation to estimate μ_i.

In estimating the small area means, Fay and Herriot (1979) used the variance stabilizing log-transformation primarily to extend the empirical Bayes estimation method proposed earlier by Efron and Morris (1975) who assumed known sampling variances for the estimates. One potential problem with such a method is the possible bias that may incur from the back-transformation. The validity of the back-transformation relies on the Taylor series argument, which may be a problem for many small area estimation problems. To avoid the problem associated with the back-transformation for the Fay-Herriot model, Chen (2001) used properties of log-normal distribution in obtaining the Bayes and empirical Bayes estimator of μ_i directly. An alternative joint modeling and estimation approach that does not require any variance stabilizing transformation is given in Liu et al. (2007).

There are now many applications of GVF in small area estimation. The U.S. Census Bureau used GVF in the Small Area Income and Poverty (SAIPE) program using the ideas given in Otto and Bell (1995). Since then a number of papers on GVF related research for poverty estimation have been written. See, e.g., Bell (2008) and Maples et al. (2009). Malec and Maples (2008) presented a related design effect formula in the context of estimating coverage error in the U.S. decennial census. Fisher (2005) and Bauder et al. (2008) proposed different GVF functions in the context of Small Area Health Insurance Estimates (SAHIE) program of the Census Bureau. At the U.S. Bureau of Labor Statistics, different GVF methods were studied in the context of Consumer Expenditure Survey and Current Employment Statistics Survey. See Cho et al. (2002), Eltinge et al. (2002), Huff et al. (2002), and Hinrichs (2003).

The choice of a reasonable working GVF model and the estimation of the parameters of the assumed GVF model are much more difficult for the small areas than for large areas. GVF models are often intuitively proposed with small area specific random effects and then the GVF model parameters are estimated using design-based variance estimates for the small areas. Because of the small area specific random effects, such modeling seems more reasonable than the GVF model with a synthetic assumption on the GVF parameters such as the slope considered in Fay and Herriot (1979). But, the small area data, especially the design-based variance estimates, could be very unreliable and noisy making it difficult to identify a reasonable GVF model with small area specific random effects. Maples et al. (2009) encountered a problem in estimating small area specific degrees of freedom parameters, resulting from Fay's successive difference replication variance method. Thus, despite the obvious shortcoming of synthetic assumptions of the Fay and Herriot approach, it may work reasonably well if the key area specific design and weight variables that influence the variability in the survey-weighted estimates are included in the GVF model.

In an unpublished manuscript, Lahiri suggested a model-assisted approach to GVF in which the determination of a reasonable GVF for the small areas is guided by a working model at the unit level that aims to capture as much variations as possible from all sources of variations in the survey estimates, including variations from the survey design and the weighting process, which involves non-response adjustment and calibration. Development of such a model needs a very good understanding of the survey design and all the detailed steps that lead to the final estimates. The variations not included in such a model is the variation of the true values around the regression surface given in Level 2. For some special cases like the one considered in this paper, it may be possible to provide an upper bound of the model variance. If the design-based variance estimates are reliable and incorporate all sources of sampling and weight variations at a higher level, one may use those to correct for the possible overestimation bias in the GVF – this was proposed in the spirit of Fay and Herriot (1979).

In this paper, we address GVF modeling issues in the context of estimating the number of school-age (5-17 year old) children in poverty for the U.S. counties using ACS. The proposed method is meant for the counties, but for now the competing GVF models are compared using the state level data since some of the design variables that we hope to include in the GVF model have not been computed at this time. Also, currently it is difficult to compare GVF models at the county level because of the lack of robust model selection criteria for small areas. This will be a good research topic for the future.

2 A Table of Notations

The following notation is for a given small area.

U: the set of all persons in the survey population for the small area

N: number of related school-age children in U

For $k \in U$,

$$y_k = \begin{cases} 1 & \text{if person } k \text{ is a related school-age child in poverty} \\ 0 & \text{otherwise} \end{cases}$$

$Y = \sum_{k \in U} y_k$, total number of school-age children in poverty in U, the main parameter of interest

$P = \frac{Y}{N}$, proportion of school-age children in poverty in U, poverty ratio.

s: ACS sample of all persons.

s_h: ACS sample of all persons in household (HH) h

$\tilde{s}$: ACS sample of school-age children

$\tilde{s}_h$: ACS sample of school-age children in HH h

a: number of households (HH) in s

n: number of persons in $\tilde{s}$

w_k: survey weight associated with person 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 sampling variability, including variability due to design and weighting

$\hat{V} \equiv \hat{V}(\hat{Y})$: Fay's successive difference replication variance estimator(Fay and Train, 1995)

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

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

We shall use lower case letters to indicate the value of a given estimator from a given sample. For example, we shall obtain $\hat{p}$ from $\hat{P}$ for a given sample. The theoretical properties of the Fay's successive difference replication method have not been investigated thoroughly. Recently, for simple random sampling, Huang and Bell (2009) attempted to understand the sampling distribution of the Fay's variance estimator using Monte Carlo simulations.

3 Model-Assisted GVF

In this paper, we shall find a model-assisted GVF for county level estimation using ideas contained in an unpublished manuscript by Lahiri. To this end, we shall first propose the following working model that is intended to approximate the ACS sample design. For any $k, k' \in \tilde{s}$,

$$\mathbf{M}: \mathrm{Cov}_M(y_k, y_{k'}) = \begin{cases} \sigma_h^2 & \text{if school-age children } k \text{ and } k' \text{ are both in HH } h, \\ 0 & \text{otherwise.} \end{cases}$$

The above model is reasonable since all HH members share an identical poor status, implying that $\mathrm{Corr}_M(y_k, y_{k'}) = 1$. Because of the binary nature of the variable, it is reasonable to assume $\sigma_h^2 = \pi_h(1-\pi_h)$, where π_h is a superpopulation true proportion of school-age children in poverty in HH h.

Under model $\mathbf{M}$, it can be shown that

$$V_M(\hat{Y}) = \left[\sum_h \pi_h(1-\pi_h)\right] \sum_h \gamma_h \Big(\sum_{k\in\tilde{s}_h} w_k\Big)^2,$$

where

$$\gamma_h = \frac{\pi_h(1-\pi_h)}{\sum_h \pi_h(1-\pi_h)}.$$

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

$$V_M(\hat{Y}) \le a\bar{\pi}(1-\bar{\pi}) \sum_h \gamma_h \Big(\sum_{k\in\tilde{s}_h} w_k\Big)^2,$$

where $\bar{\pi} = a^{-1}\sum_h \pi_h$.

In the above upper limit, if we replace the weighted average of $(\sum_{k\in\tilde{s}_h} w_k)^2$ by an unweighted average, we get the following variance function

$$V_{\text{upper}} = \bar{\pi}(1-\bar{\pi})d,$$

where $d = \sum_h (\sum_{k\in\tilde{s}_h} w_k)^2$. The above variance function motivates the following GVF:

$$V_{\text{approx}} = P(1-P)d.$$

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 still overestimate the true design-based variance. Bell (2008) examined the sensitivity of small area inference to uncertainty about sampling error variances. His research suggests that overestimation of the sampling variances is possibly less severe than underestimation. Note that in the above GVF, equivalently written as $V_{\text{approx}} = P(1-P)n\frac{d}{n}$, $\frac{d}{n}$ can be interpreted as the design effects. In the context of design effects, Gabler et al. (1999) used similar approach for a different survey design. Liu et al. (2007) obtained a design effect formula for stratified simple random sampling by using a synthetic assumption in the true design effect formula.

For a given sample, we get

$$\hat{v}_{\text{approx}} = \hat{p}(1-\hat{p})d,$$

We would like to understand the extent of bias of V_{approx} as an estimator of the design-based variance V that incorporates all sources of sampling variability. Since V and V_{approx} are unknown, we shall use the state level Fay's successive difference replication variance estimates $\hat{v}$ for V and survey-weighted estimates $\hat{p}$ for P appearing in the expression for V_{approx}.

Figure 1 probably indicates that V_{approx} overestimates V, if the Fay's successive difference variance estimator is reliable in estimating V at the state level. Ideally, to find a suitable bias-correction factor, we may want to regress the factors $b_i = \hat{v}_i/\hat{v}_{i,\text{approx}}$, for each state i, against all ACS design factors that

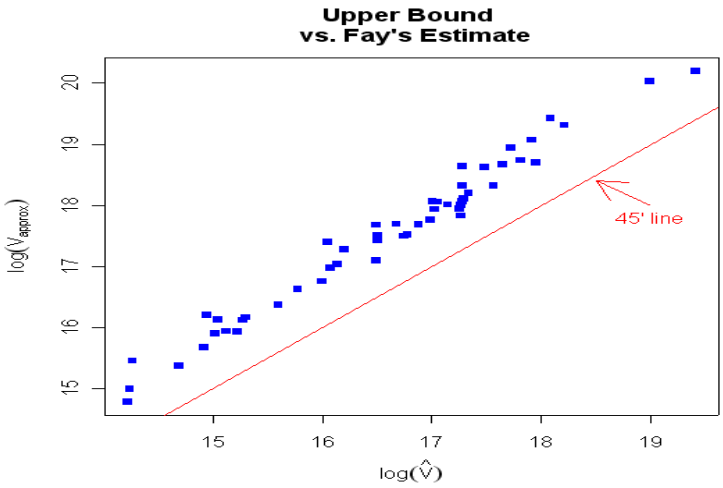

Figure 1: V_{approx} vs. Fay's estimate

were left out (e.g., response rate, CAPI rate (or rate of households that were assigned to a Computer Assisted Personal Interview), population benchmarking and sampling fraction as was done in Maples et al. (2009). In this paper, we studied the variability of the factors b_i. The next figure shows that the b_is do not vary too much across the states, especially in relation to the ACS estimate of the number of households. The graph also suggests that on the average $\hat{v}_{\text{approx}}$ is about 2.5 times of v. Based on this limited numerical work and using Bell (2008), we do not feel very uncomfortable about the possible overestimation bias of V_{approx} as an estimator of V. In any case, to correct the possible overestimation bias, we used a single adjustment factor $\bar{b} = m^{-1} \sum_{j=1}^{m} b_i$, an average of all adjustment factors for all the states. Thus one of our model- assisted GVFs is defined as $V_{MA} = \bar{b} V_{approx}$. For data analysis at the county level, we can take b_i for the state in which the county belongs. A few more model-assisted GVFs are presented in the next section.

4 Comparison of GVF Models

In this section, we shall consider five different GVF models. If the components of GVF $V_{\text{approx}} = P(1-P)d$, presented in the last section, are reasonable predictors of the true GVF, we may consider a bias-corrected GVF by fitting a multiple linear regression model (say, Model I) with $\log(\hat{v})$ as the dependent variable and the estimates of the components of $\log(\hat{V}\text{approx})$, i.e. $\log \hat{p}$, $\log(1-\hat{p})$, and $\log d$ as independent variables. Let the least squares fitted regression is given by

$$\log(\hat{v}) = b_0 + b_1 \log \hat{p} + b_2 \log(1-\hat{p}) + b_3 \log(d),$$

where b_0, b_1, b_2, and b_3 are the least squares estimates of the regression coefficients. Then the GVF motivated from Model I is given by:

$$V = \exp(b_0) P^{b_1} (1-P)^{b_2} d^{b_3}.$$

Model II is obtained from Model I by replacing $\log d$ by $\log a$. The corresponding GVF is given by:

$$V = \exp(b_0) P^{b_1} (1-P)^{b_2} a^{b_3}.$$

Model II is motivated from Maples et al. (2009).

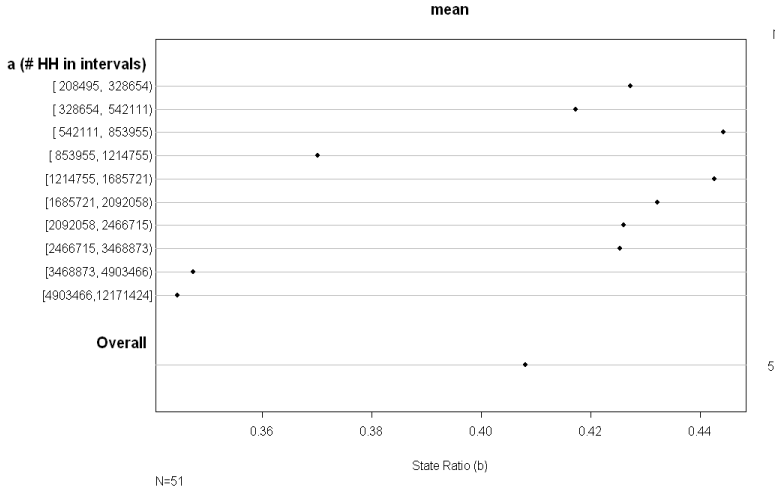

Figure 2: b VS. Number of Households

Note that $\log P$ and $\log(1-P)$ may be correlated. So in order to avoid the possible multicollinearity problem, we consider Model III where we replace the two independent variables $\log \hat{p}$ and $\log(1-\hat{p})$ by a single independent variable: $\log[\hat{p}(1-\hat{p})]$. Model III suggests the following GVF:

$$V = \exp(b_0)[P(1-P)]^{b_1} d^{b_3}.$$

Note that in the above b_0, b_1, b_2 and b_3 represent different estimates across the three models (we keep the same notation for simplicity in order to avoid new notations).

Since the above models 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 are both performing better than Model II in terms of the model selection statistics considered. Model III appears to be slightly better than Model I. Thus, the design factor d seems to be a reasonable component of the GVF.

We cannot use the model selection criteria given in Table 1 to compare GVFs motivated from Model I-III with the GVF commonly used in the CPS: $V = aY^2 + bY$ (Model IV) and $V_{\bar{b}\hat{V}_{approx}}$ (Model V) since the dependent variables are in different scales. To compare all the five models, we compute relative differences from the Fay's estimate for all the 50 states and the District of Columbia, that is $RD = (\tilde{v} - \hat{v})/\hat{v}$, where $\tilde{v}$ is one of the five variance estimates obtained from the five GVF models. Figure 4 displays the box-plots for each of the five models. Model I and Model III emerge as

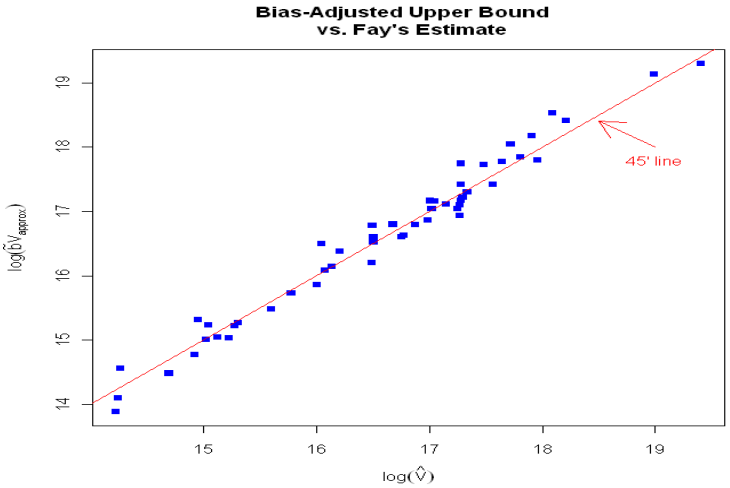

Figure 3: $\bar{b}V_{\text{approx}}$ vs. Fay's estimate

the two best performers. Model II and IV seem to have some underestimation problem (see Figure 4). The simple model-assisted estimator $\bar{b}\hat{V}_{approx}$ (Figure 3) 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. It may be possible to consider a better bias correction factor that incorporates remaining design variables.

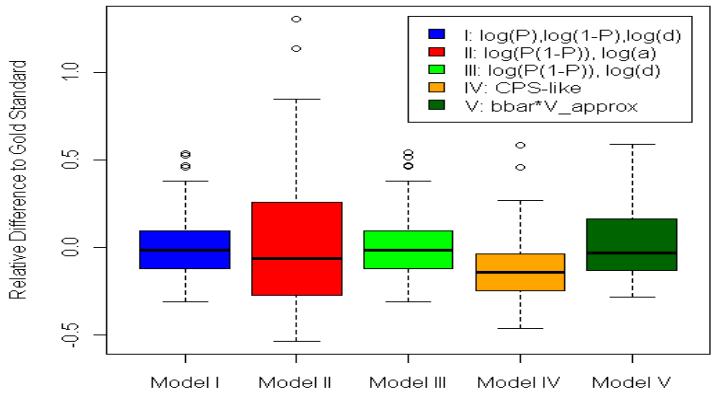

Figure 4: Comparisons based on Relative Difference

References

[1] Bauder, M., Riesz, S., and Luery, D., (2008), "Further Developments in a Hierarchical Bayes Approach to Small Area Estimation of Health Insurance Coverage: State-Level Estimates for Demographic Groups" *Proceedings of the Section on Survey Research Methods,* Alexandria, VA: American Statistical Association, pp 1726-1733.

[2] Bell, W.R. (2008), "Examining Sensitivity of Small Area Inferences to Uncertainty about Sampling Error Variances," *Proceedings of the Section on Survey Research Methods,* Alexandria, VA: American Statistical Association, pp 327-334.

[3] Chen, S. (2001), *Empirical Best Prediction and Hierarchical Bayes Methods in Small Area Estimation*, Department of Mathematics and Statistics, University of Nebraska, Lincoln.

[4] Cho, Moon, Eltinge, J., Gershunskaya, J., and Huff, L. (2002), "Evaluation of generalized variance function estimators for the U.S. Current Employment Survey," *Proceedings of the American Statistical Association, Survey Research Methods Section,* pp 534-539.

[5] Efron, B., and Morris, C. (1975), Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Asso.*. **70**, 311-9.

[6] Eltinge, J., Cho, M., and Hinrichs, P. (2002), "Use of Generalized Variance Functions in Multivariate Analysis," *Proceedings of the Section on Survey Research Methods,* Alexandria, VA: American Statistical Association, pp 904-912.

[7] Fay, R.E., and Herriot, R.Aan . (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.

[8] 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.

[9] Fisher, R. and Campbell, J. (2003), "Health Insurance Estimates for States," *Proceedings of the Government Statistics Section,* Alexandria, VA: American Statistical Association, pp 990-995.

[10] Fisher, R. and Turner, J. (2005), "Health Insurance Estimates for Counties," *Proceedings of the Section on Survey Research Methods,* Alexandria, VA: American Statistical Association, pp 1467-1473.

[11] 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.

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

[13] 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.

[14] Huang, E.T., and Bell, W. (2009), "A Simulation Study of the Distribution of Fays Successive Difference Replication Variance Estimator," *Proceedings of the American Statistical Association, Survey Research Methods Section,* pp 5294-5308.

[15] Huff, L., Eltinge, J., and Gershunskaya, J. (2002), "Exploratory Analysis of Generalized Variance Function Models for the U.S. Current Employment Survey," *Proceedings of the American Statistical Association, Survey Research Section,* pp 1519-1524.

[16] Edited IMS Lecture Notes/Monograph on Model Selection, Volume 38, 2001. Lahiri, P. ed.

[17] 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,* pp 3181-3186.

[18] Malec, D. and Maples, J. (2008), "Small Area Random Effects Models for Capture/Recapture Methods with Applications to Estimating Coverage Error in the U.S. Decennial Census," *Statistics and Medicine,* 27, 4038-4056.

[19] 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, pp 5056-5067.

[20] 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,* pp 160-165.

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

[22] U.S. Census Bureau, (2006), Design and Methodology: American Community Survey, U.S. Government Printing Office, Washington, DC. *BOC 2006* .

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