Estimating Design Effects in Small Areas/Domains through Aggregation*

Jerry J. Maples†

Abstract

Estimates of direct survey variances for areas with small sample sizes is problematic especially when combined with complex survey designs. However, reasonable estimates of these variances (or their associated design effects / effective sample sizes) are important when modeling data from these surveys. One important application is area-level small area models such as Fay and Herriot (1979). Other model-based uses include calibrating the likelihood to reflect the amount of uncertainty relative to a simple random sample. An ad hoc practice to estimate design effects or effective sample size for counts and rates in small areas and domains is to estimate the design effect for a larger aggregate and assume that the component areas or domains have the same design effect. This is valid only under restrictive conditions. Starting from the framework of a stratified sample, we will explore how the design effect at the aggregated level compares to the design effects at the lower level. We will study these design effects and propose a method to estimate them under several scenarios: unequal probability selection, unequal area means, and clustering within area.

1. Introduction

Typically, surveys are designed to contain sufficient sample size to obtain precise estimates of the quantities of interest and stable estimates of their design-based variance. However, there are many reasons why this may not be the case in practice: unplanned domains/subgroups, underestimation of nonresponse, limitations in budget to implement desired sample size, etc. When the sample size is not large enough, the estimated design-based variances can be unstable. This will impact design-based inference but can also affect model-based uses of the survey data such as area- level models for Small Area Estimation (SAE). Two types of area-level small area models are the Fay-Herriot (1979) and the Binomial Logit Normal (Franco and Bell, 2015).

In the Fay-Herriot model:

$$\begin{aligned} \hat{y}_i &= Y_i + e_i, \qquad e_i \sim N(0, V_i) \\ Y_i &= X_i^T \beta + u_i, \qquad u_i \sim N(0, \sigma^2) \end{aligned}$$

where $\hat{y}_i$ the design-based estimate of the true quantity Y_i, and V_i is the design-based variance of $\hat{y}_i$. Y_i is modeled using a regression with covariates X_i and model error u_i which has variance σ^2. The Binomial Logit Normal model is defined as:

$$\begin{aligned} y_i^* &\sim Bin(n_i^*, p_i) \\ \text{logit}(p_i) &= X_i^T \beta + u_i, \qquad u_i \sim N(0, \sigma^2) \end{aligned}$$

where n_i^* is the effective sample size, $y_i^* = \hat{p}_i n_i^*$ is the modified count scaled to the effective sample size. The effective sample size is meant to reflect the amount of information in the data after taking into account the complex sample design. The effective sample size is usually solved from the estimated sampling variance $\hat{V}_i$, i.e. $n_i^* = \hat{p}_i(1-\hat{p}_i)/\hat{V}_i$. Rounding rules may be used to keep n_i^* and y_i^* as integers. In these models, the V_i and n_i^* are treated

*Disclaimer: This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical and methodological issues are those of the author and not necessarily those of the U.S. Census Bureau.

†U.S. Census Bureau, CSRM, 4600 Silver Hill Road, Washington, D.C. 20233

as known even though they are based on the estimated design-based variance. In SAE this can have several impacts in both modeling and model assessment. In modeling, parameter estimation and prediction can be adversely affected from an unstable estimate of V_i or n_i^* by giving too much weight in the score equations to one area and too much weight toward the survey estimate in the prediction estimator. In model assessment, we can have inaccurate comparisons of model-based MSE to design-based variance estimate and distorted model diagnostic plots.

Bell (2008) conducted a sensitivity analysis of the Fay-Herriot model when the sampling variance was misspecified. The results indicated that the small area prediction estimates behaved well as long as the estimated sampling error variances were reasonably accurate. This motivates to modeling the sampling variances which should smooth out any extreme direct estimates.

2. Generalized Variance Functions (GVF)

Generalized Variance Functions (GVFs) have been used to approximate design-based variances in official surveys from the U.S. Census Bureau for many years. The most common method for GVFs is the Two-Parameter Model (Tomlin, 1974) which fits the following GVF to the design-based variance estimate of a total $\hat{x}_i$:

$$\hat{V}(\hat{x}_i) = a\hat{x}_i^2 + b\hat{x}_i \tag{1}$$

Using linearization of the rate $\hat{p} = \hat{x}/\hat{N}$ where $\hat{x}$ is a subset of $\hat{N}$ and assuming that both $\hat{x}$ and $\hat{N}$ follow (1) with the same (a, b) GVF parameters, Tomlin (1974) showed that this implies the following GVF for the rate estimate

$$\hat{V}(\hat{p}_i) = b \times \frac{\hat{p}_i(1-\hat{p}_i)}{\hat{N}_i} \tag{2}$$

where $\hat{p}_i = \hat{x}_i/\hat{N}_i$, $\hat{N}_i$ is the estimated number of people, and $\hat{N}_i/b$ is the effective sample size based on the survey design. Recent reports still use this formulation for the GVF, for example see U.S. Census Bureau (2015).

Another method to create GVFs for rate estimates is to fit a function based on the form of the variance for the binomial distribution.

$$\log\left(\hat{V}(\hat{p}_i)\right) = b_0 + b_1 \log(\hat{p}_i(1-\hat{p}_i)) + b_2 \log(n_i) \tag{3}$$

where n_i is the unweighted sample size for area i and $\hat{p}_i$ be the design-based rate. Variants of this model were fitted in Hawala and Lahiri (2010). Their method assumes a constant design effect across all areas. We propose a more direct approach to estimating the effective sample size or equivalently the design effect, using known information about the survey design (survey weights, clustering) within each area when possible and using an aggregation of areas to estimate any remaining design effects.

3. Design Effects

Let $i = 1, ..., I$ denote the area and $j = 1, ..., n_i$ denote the respondent within area i. The basic weighted survey estimator for the population rate $p_i = \sum_j y_{ij}/N_i$ is

$$\hat{p}_i = \frac{\sum_j w_{ij} y_{ij}}{\sum_j w_{ij}} \tag{4}$$

where y_{ij} is the binary indicator of some characteristic for respondent j in area i and w_{ij} is the corresponding survey weight. Our goal is to partition the sampling variance of $\hat{p}_i$ into the usual binomial variance $p_i(1-p_i)/n_i$ and multiplicative design effects which account for unequal sampling weights, clustering within the area and a shared residual design effect that is constant across all of the areas.

3.1 Unequal Weighting

Suppose the sample within an area was a random sample with varying probabilities of selection π_i, i.e. the y_is were drawn independently. The survey weights are $w_{ij} = 1/\pi_i$ and the variance of the rate is

$$Var(\hat{p}_i) = Var\left(\frac{\sum_j w_{ij}y_{ij}}{\sum_j w_{ij}}\right) = \frac{\sum_j w_{ij}^2}{(\sum_j w_{ij})^2}\sigma_{y_i}^2 = \frac{\sigma_{y_i}^2}{n_i}d_i^w \quad (5)$$

$$d_i^w = \frac{n_i \sum_j w_{ij}^2}{(\sum_j w_{ij})^2} = (1 + CV_w^2) \quad (6)$$

where $\sigma_{y_i}^2 = p_i(1-p_i)$, CV_w^2 is the squared coefficient of variation for the survey weights and d_i^w is the design effect due to unequal weighting. When all of the weights within an area are equal, i.e. sample design is simple random sampling, $d_i^w = 1$.

3.2 Clustering

Suppose area i is composed of a group of clusters $c = 1, .., C_i$ with each cluster containing n_{ic} sampled units. Suppose the design is a two-stage design: clusters are sampled (independently) and units within clusters are sampled independently, perhaps with unequal weighting (as in Section 3.1). Let y_{icj} and w_{icj} denote the observed response and associated survey weight for the j^{th} unit in cluster c for area i. From Kish (1987)

$$\begin{aligned} Var\left(\frac{\sum_{cj} w_{cj}y_{cj}}{\sum_{cj} w_{cj}}\right) &= \frac{\sigma_y^2}{n_i} \times d_i^w \left[1 + \rho_i \frac{\sum_c(\sum_j w_{icj})^2 - \sum_{cj} w_{icj}^2}{\sum_{cj} w_{icj}^2}\right] \\ &= \frac{\sigma_{y_i}^2}{n_i} \times d_i^w d_i^c \end{aligned} \quad (7)$$

where ρ_i is the intraclass correlation within clusters for area i and d_i^c is the design effect due to clustering. If there is sufficient sample size in each area then it could be possible to estimate a separate ρ_i for each area. In the presence of small sample sizes, there may be need to assumed that ρ is constant across a group (or all) of areas. The product $d_i^w d_i^c$ is the entire design effect under clustering with unequal weights.

To estimate the the intraclass correlation ρ, we use an ANOVA-like method. This is motivated by a model where the data y has a cluster-level random effect. Several estimation methods of the intraclass correlation for unbalanced binary data are given in Wu et. al. (2012) and Ridout et. al. (1999), however, all of these methods must be modified for unequally weighted data. The ANOVA-like method estimates the intraclass correlation by relating the observed sum of squares for within and between cluster means to their expected mean sums of squares. Let a $+$ in an index denote summation over that index, e.g. $w_{i++} = \sum_{cj} w_{icj}$ and let a dot in an index denote the weighted average, e.g. $\bar{y}_{ic\cdot} =$

$\sum_j w_{icj} y_{icj} / \sum_j w_{icj}$. The pooled estimate for ρ is

$$\begin{aligned}
\text{Within} \quad WSSW &= \sum_{icj} w_{icj}(y_{icj} - \bar{y}_{ic\cdot\cdot})^2 \\
MSSW &= WSSW/(w_{+++} - \sum_{ic} \frac{v_{ic}}{w_{i++}}) \qquad (8) \\
\text{Between} \quad WSSB &= \sum_{ic} w_{ic+}(\bar{y}_{ic\cdot} - \bar{y}_{i\cdot\cdot})^2 \\
MSSB &= WSSB/\left(\sum_{ic} v_{ic}(\frac{1}{w_{ic+}} - \frac{1}{w_{i++}})\right) \qquad (9) \\
\hat{\rho} &= \frac{MSSB - MSSW}{MSSB + (\bar{n} - 1)MSSW} \qquad (10)
\end{aligned}$$

where

$$\bar{n} = \frac{\sum_i (w_{i++} - \sum_c w_{ic+}^2 / w_{i++})}{\sum_{ic} v_{ic}(\frac{1}{w_{ic+}} - \frac{1}{w_{i++}})} \qquad (11)$$

and $v_{ic} = \sum_j w_{icj}^2$. If there is sufficient data in an area, i.e. large number of clusters and enough observations, one can drop the summation over areas i in (8) and (9) to estimate ρ_i. Note that even if the intraclass correlation ρ is the same across all areas, the design effect d_i^c can be different for each area due to also being a function of the survey weights within area i.

4. Pseudo-Stratified Framework

Most survey designs, such as the American Community Survey (ACS), Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP), have complex survey features beyond just unequal weighting and clustering. Design effects that we can explicitly account from the survey design should be directly estimated. In additional to traditional design effects, there are also adjustments to the survey weights for nonreponse and calibration. These adjustment procedures cause the survey weights to lose their interpretation of being the inverse probability of selection and these modification of weights also affect the estimated sampling variances. We will call this multiplicative adjustment to the design based variance from these additional unaccounted design effects and effects due to weight adjustments as the *residual design effect*.

Due to the small sample sizes, the residual design effect can not be reliably estimated for each area. Instead, we propose grouping areas together and estimating an average residual design effect treating the individual areas as if they were strata. By grouping similar areas together, the estimates of the rate and design-based variance of the rate of the aggregate will be stable.

In a stratified design, the samples from each strata are assumed to independent from each other. In some surveys, this assumption may hold when treating areas as strata. For example, counties in the ACS which are large enough to form their own group for weight adjustment are design independent. However, some counties are grouped together for weight adjustment and this would violate the independence assumption if more than one of these counties were aggregated together.

In the pseudo-stratified framework, we treat the areas as equivalent to strata. The advantage to this approach is that we can relate the variance of the individual areas explicitly to the variance of the aggregate, which can be stably estimated. Let A denote the set of areas that will be aggregated, N_i is the population size of area i and $N = \sum_{i \in A} N_i$. The aggregate rate estimate and its variance under a stratified design is:

$$\hat{p} = \sum_{i \in A} \frac{N_i}{N} \hat{p}_i = \sum_{i \in A} \frac{N_i}{N} \frac{\sum_j w_{ij} y_{ij}}{\sum_j w_{ij}} \tag{12}$$

$$Var(\hat{p}) = \sum_{i \in A} \left(\frac{N_i}{N} \right)^2 \frac{p_i(1-p_i)}{n_i} d_i^w d_i^c d_i \tag{13}$$

where d_i is the residual design effect. The term d_i represents the correction factor needed for the sampling variance to fully account for all of the design features of the survey (including all of the weight adjustment procedures such as nonresponse and calibration) that are not explicitly represented. Two big additional assumptions will be made here. First is that $\sum_j w_{ij} \approx N_i$ and second is that the d_i's are all equal, i.e. $d_i = d$. The first assumption only holds if that subpopulation matches a calibration cell for the survey (or approximately true for a very large sample in a design consistent survey). The second assumption can also be viewed as estimating an average design effect given that we don't have the data to estimating each individual d_i. While these assumptions may seem overly strong, results from Bell (2008) showed that the estimated sampling variances do not need to be highly accurate (say only within a factor of 2) for shrinkage style estimates of the means to maintain good properties. One of the largest concerns in using unstable sampling variance estimates is when an estimated variance is much smaller than its true value which makes the estimate appear to be very precise. This affects both parameter estimation and prediction.

Putting these assumptions together we can rewrite the estimated rate and variance of the estimated rate for the aggregate of areas as

$$\hat{p} = \frac{\sum_{i \in A} \sum_j w_{ij} y_{ij}}{\sum_{i \in A} \sum_j w_{ij}} \tag{14}$$

$$Var(\hat{p}) = \sum_{i \in A} \left(\frac{N_i}{N} \right)^2 \frac{p_i(1-p_i)}{n_i} d_i^w d_i^c d_i \tag{15}$$

$$= \sum_{i \in A} \frac{\sum_j w_{ij}^2}{(\sum_{i \in A} \sum_j w_{ij})^2} \frac{p_i(1-p_i)}{n_i} d_i^c d \tag{16}$$

Note that if the p_is and d_i^cs are all constant within the aggregate then (16) reduces to estimating the residual design effect for the aggregate after taking unequal weighting into account. The ad hoc method of using the design effect of the aggregate is only valid when the true rate and design effects (after unequal weighting is taken into consideration) are equal across all of the areas. The solution for d in (16) is

$$\hat{d} = Var(\hat{p}) \left[\sum_{i \in A} \frac{\sum_j w_{ij}^2}{(\sum_{i \in A} \sum_j w_{ij})^2} \frac{p_i(1-p_i)}{n_i} d_i^c \right]^{-1} \tag{17}$$

In order to use (17), we need to provide the area level rates, p_i. The survey estimates of p_i are usually too unstable to use. Some areas may have a large enough sample to provide good estimates of p_i but we expect that many areas will not. If the rates between the areas do not vary much, then substituting the aggregate rate $\hat{p}$ for p_i will approximate (17) reasonably well. Two additional methods for estimating p_i for the GVF in (17) are proposed.

4.1 Model-Based Estimator for p_i

The first method is to fit a separate model for the p_i's. This is plausible in the context of small area estimation when there is already a model for the rate and a set of covariates in mind. Instead of using the full small area model to estimate p_i, a simpler version is suggested. A preliminary model-based estimate of p_i, say p_i^m, can be obtained by fitting a weighted least squares to the mean function of the small area model. Suppose the mean function is $\mu(X_i^T\beta)$, where $\mu(\cdot)$ is the inverse logistic function. Then the parameters β can be estimated by minimizing the objective function $\sum_{i\in A} n_i(\hat{p}_i - \mu(X_i^T\beta))^2$ and $p_i^m = \mu(X_I^T\hat{\beta})$. This approach was used in Carolina and Bell (2015). Under this modeling framework, using the aggregate estimated rate $\hat{p}$ is the same as having an intercept only model for the mean. After the model-based p_i^m are obtained, the residual design effect is estimated using (17).

4.2 Shrinkage-Based Estimator for p_i

An alternative method when there is no assumed model for p_i or covariates to predict is to create a shrinkage estimator. This second method creates a composite estimator for p_i from $\hat{p}$ and $\hat{p}_i$ which minimizes the mean squared prediction error based on moment assumptions. We use a simple underlying superpopulation model: $p_i = p_{sp} + u_i$ where $Var(u_i) = \sigma_p^2$ and p_{sp} is the superpopulation mean of the area rates. Assume the summation over areas (subscript i) are for those in aggregate A. Let $d_i^* = d_i^w d_i^c d$ and $\gamma_i = N_i/N$, then $\hat{p} = \sum_i \gamma_i \hat{p}_i$ and $p = \sum_i \gamma_i p_i$. From the superpopulation model framework:

$$Var(\hat{p}_i|p_i) = \frac{p_i(1-p_i)}{n_i} d_i^* \tag{18}$$

$$Var(\hat{p}_i) = \frac{p_i(1-p_i)}{n_i} d_i^* + \sigma_p^2 \tag{19}$$

$$Var(\hat{p}) = \sum_i \gamma_i^2 \frac{p_i(1-p_i)}{n_i} d_i^* + \sigma_p^2 \sum_i \gamma_i^2 \tag{20}$$

Using method of moments, for known d_i^* and p_i's, the estimator for σ_p^2 is:

$$E\left[\sum_i \gamma_i(\hat{p}_i - \hat{p})^2\right] = E\left[\sum_i \gamma_i \left(((\hat{p}_i - p_{sp}) - \sum_k \gamma_k(\hat{p}_k - p_{sp})\right)^2\right]$$

$$= \sum_i \gamma_i(1-\gamma_i)p_i(1-p_i)d_i^*/n_i + (1 - \sum_i \gamma_i^2)\sigma_p^2 \tag{21}$$

$$\Rightarrow \hat{\sigma}_p^2 = \frac{\sum_i \gamma_i(\hat{p}_i - \hat{p})^2 - \sum_i \gamma_i(1-\gamma_i)p_i(1-p_i)d_i^*/n_i}{1 - \sum_i \gamma_i^2} \tag{22}$$

It is possible for the solution of σ_p^2 to be negative in (22). In this case, one can set $\sigma_p^2 = 0$ which indicates that the variability in the p_is is not enough to overcome the uncertainty from the sampling design. The composite estimator for p_i has the form of $\alpha_i\hat{p}_i + (1-\alpha_i)\hat{p}$. Areas with lots of data, i.e. large effective sample size, should have an estimate closer to their $\hat{p}_i$ and areas only few sample data should have an estimate closer to $\hat{p}$. To solve for

α_i, we minimized the expected mean squared prediction error

$$\begin{aligned}
\min_{\alpha_i} E\Big[\alpha_i\hat{p}_i + (1-\alpha_i)\hat{p} - p_i\Big]^2 &= \min_{\alpha_i} E\Big[\alpha_i(\hat{p}_i - p_i) + (1-\alpha_i)(\sum_k \gamma_k\hat{p}_k - p_i)\Big]^2 \\
&= \min_{\alpha_i} E\Big[(\alpha_i + (1-\alpha_i)\gamma_i)(\hat{p}_i - p_i) + (1-\alpha_i)\sum_{k\neq i}\gamma_k(\hat{p}_k - p_i)\Big]^2 \\
&= \min_{\alpha_i} \Big[\big(\alpha_i^2 + 2\alpha_i(1-\alpha_i)\gamma_i\big)p_i(1-p_i)d_i^*/n_i \\
&\quad + (1-\alpha_i)^2\sum_k \gamma_k^2 p_k(1-p_k)d_k^*/n_k \\
&\quad + (1-\alpha_i)^2\sigma_p^2\big(1 - 2\gamma_i + \sum_k \gamma_k^2\big)\Big] \qquad (23)\\
\Rightarrow 0 &= (\alpha_i(1-2\gamma_i) + \gamma_i)p_i(1-p_i)d_i^*/n_i - (1-\alpha_i)\sum_k \gamma_k^2 p_k(1-p_k)d_k^*/n_k \\
&\quad - (1-\alpha_i)\sigma_p^2(1 - 2\gamma_i + \sum_k \gamma_k^2) \qquad (24)\\
\Rightarrow \alpha_i &= \frac{\sum_k \gamma_k^2 p_k(1-p_k)d_k^*/n_k - \gamma_i p_i(1-p_i)d_i^*/n_i + \sigma_p^2(1-2\gamma_i + \sum_k\gamma_k^2)}{(1-2\gamma_i)p_i(1-p_i)d_i^*/n_i + \sum_k\gamma_k^2 p_k(1-p_k)d_k^*/n_k + \sigma_p^2(1-2\gamma_i+\sum_k\gamma_k^2)} \qquad (25)\\
\Rightarrow \alpha_i &\approx \frac{p(1-p)\sum_k\gamma_k^2 d_k^*/n_k - \gamma_i p(1-p)d_i^*/n_i + \sigma_p^2(1-2\gamma_i+\sum_k\gamma_k^2)}{(1-2\gamma_i)p(1-p)d/n_i^* + p(1-p)\sum_k\gamma_k^2 d_k^*/n_k + \sigma_p^2(1-2\gamma_i+\sum_k\gamma_k^2)} \qquad (26)
\end{aligned}$$

Since the estimation of d, σ_p^2 and p_i all depend on each other, one can start off with initial guess of p_i as the aggregate rate $\hat{p}$ and then iterate between the equations. However, in some data examples, several of the α_is were converging to 1 when $\hat{p}_i = 0$ and this is not a desirable solution since most of the areas that have $\hat{p}_i = 0$ also have a very small sample size. For the solution of α_i, it may be better to use the approximation in (26) which smooths over all of the individual p_i's using $\hat{p}$. Given the solution to α_i we can construct our composite estimate of p_i

$$\tilde{p}_i = \alpha_i\hat{p}_i + (1-\alpha_i)\hat{p} \qquad (27)$$

4.3 GVF and Effected Sample Size Estimator

With the estimated of the design effect $\hat{d}_i^* = d_i^w d_i^c \hat{d}$ and an estimate for p_i, we can construct a GVF based sampling variance estimator by

$$\hat{V}(\hat{p}_i) = \tilde{p}_i(1-\tilde{p}_i)\hat{d}_i^*/n_i \qquad (28)$$

Note that if there is no clustering design effect (or if it was not estimated) then $d_i^c = 1$. The effect sample size is $n_i^* = n_i/d_i^*$. This GVF can be used as an estimate of the design-based variance for $\hat{p}_i$ and the GVF and effective sample size can be used as inputs into small area models.

The GVF in (28) will be consistent with the sampling variance of the aggregate provided that the assumptions of the pseudo-stratified framework hold. The method outlined in this section are to create reasonable estimates of the sampling variance when the data does not support using more traditional GVF modeling approaches.

5. Applications

5.1 Tract Level Child Poverty Rates

The American Community Survey (ACS) is a large scale general survey designed to make estimates at various geographic levels, e.g. states and counties. The ACS is a cross sectional survey drawing a new sample each year. The ACS releases 1-year estimates for all states and counties that contain more than 65000 in total population and 5-year estimates for all areas (states, counties and tracts). We are interested in poverty rates for school-aged children (age 5-17) in tracts. For demonstration of the method, we will focus on the 5-year ACS (2015) data from Charles County, Maryland which contains 30 tracts and 1983 school-aged children in sample. The sample sizes ranging from 30-111 children per tract with 1983 school-aged children in total.

The overall school-age child poverty rate for the county was 9.6%, with 3 of the 30 tracts having a direct estimate of 0%. Poverty data is also clustered by definition because the poverty indicator is a household level variable and all persons in the same household have the same poverty status. We will consider census blocks as the sub-tract unit for clustering, which will incorporate the local area effect and within household effect. There were 534 census blocks distributed across the 30 tracts. Many of the census blocks only had a single household with school-aged children in sample. Due to the small sample sizes in most tracts, we used the pooled estimate of intraclass correlation across all tracts and estimated ρ to be .41. There were no covariates to estimate the tract level child poverty rate, therefore the shrinkage estimate for p_i was used.

Two GVF models were fitted. The first model, GVF(w), only took the unequal weighting directly into account. Therefore the clustering effect has to be absorbed into the estimate of $\hat{d}$. The second model, GVF(wc), included both the design effect for unequal weighting and the estimated cluster effect for each area, using the pooled intraclass correlation.

$$\text{GVF(w):} \quad \hat{V}(\hat{p}_i) = \tilde{p}_i(1-\tilde{p}_i)d_i^w \hat{d}/n_i \tag{29}$$
$$\text{GVF(wc):} \quad \hat{V}(\hat{p}_i) = \tilde{p}_i(1-\tilde{p}_i)d_i^w \hat{d}^\dagger/n_i \tag{30}$$

Table 1 shows the range of the design effects due to unequal weighting and clustering and the product of two design effects. In addition, it shows the estimated residual design effect, $\hat{d}$ which was estimated by the method given in Section 4. When clustering is not taken into account, a residual design effect of 2.84 was needed so that the tracts sampling variance matched the direct survey variance of the county assuming a stratified design. We expected a significant residual design effect due to the large estimate of intraclass correlation and knowing that poverty is perfectly correlated within a household. When the clustering at the census block level was explicitly taken into account, the residual design effect $\hat{d}^\dagger$ was estimated at .82. Thus, adding in the clustering effect overcompensated so the residual design effect had to raise the effective sample size to keep consistency with the county direct variance estimate. The tracts within the same county are all part of the same adjustment group for calibration to known population totals. This calibration can add dependency by making the areas positively correlated due to common adjustment factor, thereby violating the stratified framework assumption.

Figure 1 shows the comparison of the square root of the direct sampling variance at the tract level compared to the square root of the GVF estimate for the sampling variance under the GVF model with and without the explicit clustering design effect. The GVFs greatly reduces the range of the sampling variance across the tract, perhaps over smoothing. The three tracts that had direct estimates of 0 (and corresponding sampling variances of 0) have

Design Effect	Min	Median	Max	Residual Design Effect	$\hat{\sigma^2}_p$
d_i^w	1.16	1.41	1.83	$\hat{d}$= 2.84	.00606
d_i^c	1.63	2.62	5.31		
$d_i^w \times d_i^c$	2.31	3.75	8.76	$\hat{d}^\dagger$ =.82	.00636

Table 1: Range of design effects for unequal weights d_i^w, clustering d_i^c and the product $d_i^w \times d_i^c$ for tract-level poverty rates of school-aged children from the ACS

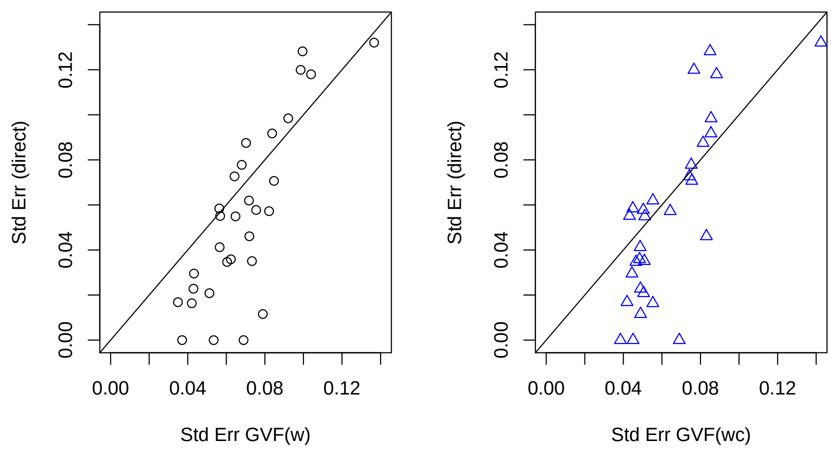

Figure 1: Comparison of Direct Sampling Variance (y-axis) and GVFs (x-axis). GVF(w) [left plot] uses unequal weighting design effect and GVF(wc) [right plot] uses unequal weighting and clustering design effects.

GVF estimates well away from 0, allowing for a better estimate of uncertainty about the estimated rate.

5.2 State Level Disability Rates

The Survey of Income and Program Participation (SIPP) is a nationally representative sample to estimate the sources and amounts of income, labor force information, program participation and eligibility data, and general demographic characteristics. The SIPP survey is a longitudinal survey which interviews the sampled respondents every 4 months, called a wave. The survey contains a set of core questions (asked at every wave) and sets of topical modules which are only asked during specific waves, e.g. child care, wealth, disability, and school enrollment. The disability module has detailed questions (the same questions as in the standard Activities of Daily Living and Instrumented Activities of Daily Living Battery) about many different aspects of disability. We are interested in estimating the rate of people aged 15 and above that have any disability for each state.

Starting with the 2008 sample, the ACS added six general questions on disability, covering: hearing, vision, cognitive, ambulatory, self-care and independent living difficulties. Many of the national and state level estimates of disability are higher from the SIPP due to the more comprehensive set of questions about disability that are asked than from the ACS.

Design Effect	Min	Median	Max	Residual Design Effect
d_i^w	1.08	1.18	1.28	$\hat{d}$ = 1.55
d_i^c	1.50	3.48	17.68	
$d_i^w \times d_i^c$	1.75	3.88	20.12	$\hat{d}^\dagger$ = .24

Table 2: Range of design effects for unequal weights d_i^w, clustering d_i^c and the product $d_i^w \times d_i^c$ for state-level disability rates of people aged 15+ from SIPP

However, the ACS estimates of disability are based on a large sample size and thus very stable at the state level and could be used as a predictor for the SIPP disability rate. We create a preliminary model for the SIPP disability rate and solve for $\beta = (\beta_0, \beta_1)$ minimizing the following weighted least squares function:

$$\sum_i n_i(\hat{p}_{i,SIPP} - \text{logit}^{-1}(\beta_0 + \beta_1 \text{logit}(X_{i,ACS})))^2 \quad (31)$$

where logit^{-1} is the inverse logit transformation and $X_{i,ACS}$ is the state level disability rate estimated by the ACS. The fitted values for p_i were $p_i^m = \text{logit}^{-1}(.353 + .942\text{logit}(X_{i,ACS}))$.

The two GVF models, GVF(w) and GVF(wc), are the same as given in (29) and (30). Table 2 shows the range of design effects and the estimated residual design effect for the state level disability estimates. Since counties were randomly selected within state in the survey design, we used the counties for the clustering group within state. The intraclass correlation was small, $\rho = .024$. However, the design effects due to clustering had some large values. The cluster design effect can get very large when the cluster sizes are large and the number of clusters is small. The residual design effect for the GVF(wc) had to compensate for the large cluster design effects. The model GVF(w) only needed a moderate adjustment with the residual design effect of 1.55.

Figure 2 shows the comparisons of the direct sampling variances from the SIPP state level disability estimates to the two GVF model estimates. Since the sample sizes in this SIPP example are larger than the previous ACS example (average sample size per area is 20 times larger), the comparison of the GVFs to the direct sampling variance estimate gives insight into whether we should include the clustering design effect. The plot without the clustering effect (on left) shows a more even spread of the direct sampling variance around the GVF. The plot with the clustering effect (on right) is underestimating the sampling variance when the SIPP sampling variance is large (states with smaller sample size). For this dataset, given the large cluster design effects with the small intraclass correlation and visual evidence from the plots, we would not recommend using the cluster design effect in the construction of the GVF.

6. Discussion

A generalized variance function (GVF) was developed for rate estimates from complex survey data for area with small sample sizes by using layers of design effects. Some of the design effects explicitly take into account features of the survey design such as unequal weighting and clustering. A residual design effect which captures the unaccounted part of the overall design effect for an area was estimated by the average of the residual design effect over multiple similar areas. The areas were linked together by using a stratified design where the areas were treated as strata, which was called a pseudo-stratified framework. Unlike estimating traditional GVFs, the area sampling variances were never directly used

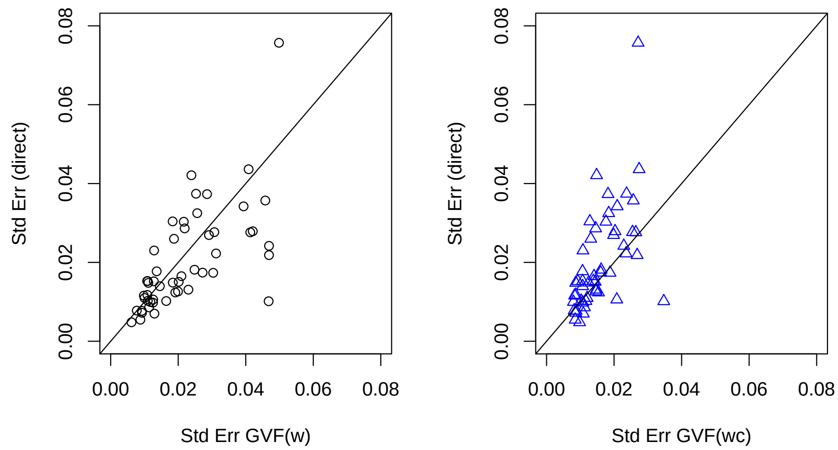

Figure 2: Comparison of Direct Sampling Variance (y-axis) and GVFs (x-axis). GVF(w) [left plot] uses unequal weighting design effect and GVF(wc) [right plot] uses unequal weighting and clustering design effects.

to fit any of the parameters. If many areas have sufficient sample size and only a few were too small then the traditional GVF methods would be valid. This procedure was developed for cases where a large portion of the areas had sampling variance estimates that were too noisy to even fit GVFs, e.g. where many areas had direct rate estimates of 0 or 1.

Two methods were given for what to use for the estimated area rate p_i instead of the direct survey estimate $\hat{p}_i$ which may be poorly estimated due to small sample size. The first method depended on a mean model and auxiliary covariates. If the goal is small area modeling, then these will already be available. A second method was developed when there was no model or covariate to use to estimate p_i which was based on a shrinkage estimator between the aggregate rate and the area rate which minimized the expected mean squared prediction error.

How well this pseudo-stratified framework approach works depends on the underlying design independence assumption of the areas that compose the aggregate. It also depends on the similarity of the residual design effects across the areas. Care should be taken on how to chose the aggregates. They don't always need to be within existing nested geographies but if the areas are similar in both their true rates and residual design effects then the approximations that average over the areas should work well.

REFERENCES

Bell, W. (2008). "Examining Sensitivity of Small Area Inferences to Uncertainty about Sampling Error Variances," *JSM Proceedings, Survey Methods Section*, Alexandria, VA: American Statistical Association.

Fay, R.E. and Herriot, R.A. (1979). "Estimation of Income from 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 and States and Counties", *JSM Proceedings, Government Statistics Section*, Alexandria, VA: American Statistical Association, pp 154-159.

Franco, C. and Bell, W. (2015). "Borrowing Information over Time in Binomial/Logit Normal Models for Small Area Estimation," joint issue of *Statistics in Transition* and *Survey Methodology*, **16(4)**, pp 563-584.

Hawala, S. and Lahiri, P. (2010)/ "variance Modeling in the U.S. Small Area Income and Poverty Estimates

Program," *JSM Proceedings, Survey Research Methods Section*, Alexandria, VA: American Statistical Association.

Kish, L. (1987) "Weighting in the Deft," The Survey Statistician, June 1987.

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

Ridout, M., Demetrio, G., and Firth, D. (1999) "Estimating Intraclass Correlation for Binary data," *Biometrics*, **55**, pp 137-148.

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 (2015). Source and Accuracy of Estimates for Income and Poverty in the United States: 2014 and Health Insurance Coverage in the United States: 2014,
http://www2.census.gov/library/publications/2015/demo/p60-252sa.pdf

Wu, S., Crespi, C., and Wong, W. (2012). "Comparison of Methods for Estimating the Intraclass Correlation Coefficient for Binary Responses in Cancer Prevention Clustered Randomized Trials," Contemporary Clinical Trials, 33 (5), pg 869-880.