Small Area Confidence Bounds on Small Cell Proportions in Survey Populations*

Aaron Gilary[1], Jerry Maples[1], Eric V. Slud[1,2]
[1]U.S. Census Bureau, CSRM, 4600 Silver Hill Road, Washington, DC 20233
[2]Mathematics Department, University of Maryland, College Park, MD 20742

Abstract

Motivated by the problems of 'quality filtering' of estimated counts in American Community Survey (ACS) tables, and of reporting small-domain coverage results from the Census Coverage Measurement (CCM) program, this paper studies methods for placing confidence bounds on proportions for cells and tables, estimated from complex surveys, in which the estimated counts are zeroes. While coefficients of variation are generally used in measuring the quality of estimated counts, they do not make sense for assessing validity of very small estimated counts. The problem is formulated here in terms of (upper) confidence bounds for unknown proportions. We discuss methods of creating confidence bounds from small-area models including logistic, beta-binomial, and variance-stabilized (arcsin square root transformed) linear models. The model-based confidence bounds are compared with single-cell bounds derived from arcsin-square-root transformed binomial intervals with survey weights embodied in the 'effective sample size'. The comparison is illustrated on county-level data about Housing-Unit Erroneous Enumeration status from the 2010 CCM.

Key Words: arcsin square root transformation, confidence bound, parametric bootstrap, prediction interval, beta-binomial regression, transformed Fay-Herriot model

1. Introduction

Within the same year, the Census Bureau encountered two similar problems relating to the estimation of proportions in small domains. The first one was to provide bounding intervals of small estimated proportions in American Community Survey (ACS) tabulations. The second was to construct a measure of uncertainty for county and place level estimates of Erroneous Enumeration (EE) rates among Housing Units in the Census Coverage Measurement (CCM) program.

These problems have three salient common features that define a generic problem in survey estimation. First, they require interval estimates corresponding to survey point estimates with values near zero. This requires special treatment because the straightforward design-based variance estimators of small proportions yield interval estimators unrealistically close to zero. The second common feature is the possibility of using small-area estimation techniques, both for point and interval estimators, because potentially useful covariates are available. These covariates may

*This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

be the categorical demographic variables (geography, race, age, etc.) defining a fine cross-classification, as in the tables of the ACS, or may arise from a combination of demographic variables and extensive paradata features of a well-documented frame, as in the CCM example. The third common feature is the availability of a strong synthetic predictor: as often occurs in small-domain survey problems, a local-area proportion can be roughly estimated by the corresponding proportion over an appropriately chosen higher level of aggregation, as when county rates are approximated by the analogous state-wide rates. As in other small-area survey problems, a recurring theme is the balancing of an aggregated, possibly biased data source versus a smaller, highly variable one.

This paper will discuss general methods for using small-area techniques for one-sided interval estimation of small survey proportions, and for comparing these methods to 'cell-based' or direct interval estimates that do not borrow strength across small domains or cells. The ACS application was discussed in Slud (2012) and is still under investigation. In this paper, methods and results will apply to the CCM EE rates for housing units.

The paper is organized as follows. In Section 2, the data structure, notation, models and methods are defined. Section 3 presents the necessary complication of *effective sample sizes* which are to be used as sample sizes either in small-area models or in transformed-scale confidence intervals. Section 4 provides background on the CCM program to which the techniques of the paper are applied. Section 5 describes the choice of covariates entering the small-area regression-type models. Finally, in Section 6 we compare and interpret the numerical results on the CCM data and give a preliminary assessment at the level of the larger counties of the relative value of cell-based versus small area one-sided upper-bounding interval estimates. Conclusions and remaining research questions are summarized in Section 7.

2. Setting, Models, and Methods

The setting is a data structure in which m small domains or 'areas' $i = 1, \ldots, m$ are equipped with survey-weighted (ratio) estimators $\hat{\bar{y}}_i \equiv \hat{Y}_i/\hat{N}_i$, where N_i denotes a (generally unknown) domain population size, and where the survey-estimated proportion $\hat{\bar{y}}_i$ is defined equal to the further expressions

$$\hat{\bar{y}}_i \;=\; \frac{\hat{Y}_i}{\hat{N}_i} \;=\; \frac{y_i}{n_i} \;=\; \frac{y_i^*}{n_i^*} \tag{1}$$

Here $\hat{Y}_i$ and $\hat{N}_i$ are the area y-indicator total and population-total estimates, respectively. The denominator n_i is the actual number of people or units sampled in area i, while y_i is a derived rather than an observed quantity, interpreted as the (estimated) count of y-indicators of 1 among the sample. The denominator n_i^* is a so-called *effective sample size* related to n_i/DEFF_i where DEFF denotes the 'design effect', the ratio of design-based sample variance of $y_i = n_i\,\hat{\bar{y}}_i$ to what it would be under a simple random sampling design. Alternative methods for defining the effective sample size n_i^* are discussed in Section 3 below. Then y_i^* is interpreted as a sample count associated with the effective sample size n_i^* and is defined by (1). The terms used in (1) are set so that the three ratios are exactly equal.

Despite the use of the terms y_i^* and n_i^* as counts and sample sizes, these quantities defined in Sec. 3.2 and in (1) are generally not integer values. Nevertheless, we maximize the same likelihoods defined for the models (4) – (6) to generate parameter estimates. However, when models (5) and (6) are used in generating parametric

bootstrap samples, we use 'stochastic rounding' which means that samples with non-integer sample size n_i^* are drawn from the next-larger integer sample size with probability equal to the fractional part of n_i^*, and otherwise from the next-smaller integer sample size.

We compare four methods of construction of upper confidence bounds (UCBs) for survey-based estimates of small proportions $\bar{y}_i$, building on a similar comparison of the first three which was undertaken by Slud (2012). The first, called a *Cell-based* method, is a direct method not based on a model with shared parameter across areas i. The other methods are based on small area estimation models, which take three different forms but share the feature that the unknown survey proportions have a generalized linear regression form in terms of covariates $\mathbf{x}_i$, with error or dispersion measured by a single additional unknown parameter. Each of the three models also involves a nonlinear transformation from regression scores $\mathbf{x}_i^{tr}\,\beta$ to the probability scale.

2.1 Cell-Based Method

The Cell-based method assumes a simple binomial model for the untransformed data, with an effective sample size of n_i^* and a rate $\pi_i = \bar{y}_i$ for area i. The arcsin square-root transformation is a variance-stabilizing transformation for binomial proportions, which means that for binomial or simple random sampled data, the variance of this transformation of the sample mean approximately, via the Delta Method, does not depend on the underlying rate π_i. That is, when n_i^* is moderately large,

$$y_i^* \sim \text{Binom}(n_i^*\,,\,\bar{y}_i) \implies \arcsin(\sqrt{y_i^*/n_i^*}) \overset{\mathcal{D}}{\approx} \mathcal{N}(\arcsin\sqrt{\bar{y}_i},\, \frac{1}{4n_i^*}). \qquad (2)$$

Based only on data y_i^* from area i, a standard one-sided confidence interval for $\pi_i = \bar{y}_i$ is then derived from the level $1-\alpha$ confidence statement

$$\arcsin(\sqrt{y_i^*/n_i^*}) \,-\, \arcsin(\sqrt{\bar{y}_i}) \;\geq\; -z_\alpha/\sqrt{4n_i^*}$$

which is immediately transformed back to the probability scale to give

$$\textbf{UCB.cell}_i \;=\; \min\left\{1,\, \sin^2\left(\arcsin(\sqrt{y_i^*/n_i^*}) \,+\, z_\alpha\,/\sqrt{4n_i^*}\,\right)\right\} \qquad (3)$$

Although this method is simple to describe, correct for large samples, and leads to bounds not so inflated near $\pi_i = 0$ as several of the otherwise good one-sided intervals compared by Liu and Kott (2009), it is known to have quite anticonservative (i.e., smaller than nominal) coverage for underlying proportions $\bar{y}_i = \pi_i \leq 0.25$. A simple modification which widens the interval and gives it conservative coverage is to replace y_i^*/n_i^* in (3) by $(y_i^*+1)/(n_i^*+2)$. A compromise method which we study further in a forthcoming technical report is to replace y_i^*/n_i^* in (3) by $(y_i^*+1/2)/(n_i^*+1)$.

2.2 Small Area Models and Estimators

We next consider methods of constructing UCBs based on small area models. Throughout, we assume that area-level covariates $\mathbf{x}_i$ are available and can be treated as known design constants. For each model, we specify jointly a form for both the *area-level target* $\bar{y}_i$ and the observation y_i^*/n_i^*, and in each case the target

is expressed in terms of a linear combination $\eta_i \equiv \mathbf{x}_i^{tr}\beta$ of components of $\mathbf{x}_i$ and a further unmodelled area-level random effect. First, we consider a *transformed Fay-Herriot* model based on the classic model of Fay and Herriot (1979).

$$\textbf{FHtr}: \quad \left\{ \begin{array}{rcl} \arcsin(\sqrt{y_i^*/n_i^*}) & = & \arcsin(\sqrt{\bar{y}_i}) \; + \; e_i, \qquad e_i \sim \mathcal{N}(0,\, 1/(4n_i^*)) \\ \arcsin(\sqrt{\bar{y}_i}) & = & \mathbf{x}_i^{tr}\,\beta \; + \; u_i \; , \quad u_i \sim \mathcal{N}(0, \sigma_u^2) \end{array} \right. \tag{4}$$

The point of the transformation within this model is the same as in the cell-based transformed sample mean in Section 2.1.

The other small-area models both treat y_i^* as binomial with n_i^* trials and success-probability $\bar{y}_i$, with the mean of the latter expressed in terms of the logistic distribution function $h(x) = e^x/(1+e^x)$. One of these models is the *random-intercept logistic*, discussed by Jiang and Lahiri (2006).

$$\textbf{GLMM}: \quad \left\{ \begin{array}{rcl} y_i^* & \sim & \text{Binom}(n_i^*, \bar{y}_i) \\ \bar{y}_i & = & h(\mathbf{x}_i^{tr}\,\beta + v_i) \; , \quad v_i \; \sim \; \mathbf{N}(0, \sigma_v^2) \end{array} \right. \tag{5}$$

The final model is the *beta-binomial* employing a logit link specifically studied in Prentice (1986).

$$\textbf{BBIN}: \quad \left\{ \begin{array}{l} y_i^* \; \sim \; \text{Binom}(n_i^*, \bar{y}_i) \\ \bar{y}_i \; \sim \; \text{Beta}\Big(h(\mathbf{x}_i^{tr}\,\beta) \; \cdot \; \tau, \; (1 - h(\mathbf{x}_i^{tr}\,\beta)) \; \cdot \; \tau\Big) \end{array} \right. \tag{6}$$

Within all three small area models, the regression coefficients β are unknown and are estimated jointly by maximum likelihood with the respective random-area-effect dispersion parameter σ_u^2, σ_v^2, or $1/(1+\tau)$.

In the **FHtr** model, the *empirical best* point predictor (EBP) for the target $\bar{y}_i$ is defined (Rao 2003) by substituting maximum likelihood estimators (MLEs) $(\hat{\beta},\, \hat{\sigma}_u^2)$ for the unknown parameters (β, σ_u^2) into the expression $E(\arcsin(\sqrt{\bar{y}_i}) \,|\, y_i^*)$, yielding

$$\arcsin(\sqrt{\widehat{\bar{y}}_i^{\text{BP}}}) \; = \; \mathbf{x}_i^{tr}\,\hat{\beta} \; + \; \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 \; + \; (4n_i^*)^{-1}} \; \Big(\arcsin(\sqrt{y_i^*/n_i^*}) \; - \; \mathbf{x}_i^{tr}\,\hat{\beta} \Big) \tag{7}$$

In each of the **GLMM** and **BBIN** models, the empirical best predictors for the target $\bar{y}_i$ is given by substituting MLEs for parameters in the conditional expectation $E(\bar{y}_i \,|\, y_i^*)$. The resulting expressions are, for **GLMM**:

$$\widehat{\bar{y}}_i^{\text{BP}} \; = \; g(y_i^* + 1,\, n_i^* + 1,\, \mathbf{x}_i^{tr}\hat{\beta},\, \hat{\sigma}_v^2) / g(y_i^*,\, n_i^*,\, \mathbf{x}_i^{tr}\hat{\beta},\, \hat{\sigma}_v^2) \tag{8}$$

where

$$g(k, n, \eta, \sigma^2) \; \equiv \; \int \frac{e^{(\eta + \sigma z)k}}{(1 + e^{\eta + \sigma z})^n}\, \phi(z) dz \quad , \quad \phi(\cdot) \sim \mathcal{N}(0,1) \tag{9}$$

and in the **BBIN** model,

$$\widehat{\bar{y}}_i^{\text{BP}} \; = \; \{y_i^* \; + \; \hat{\tau}\, h(\mathbf{x}_i^{tr}\hat{\beta})\} \,/\, \{n_i^* \; + \; \hat{\tau}\} \tag{10}$$

recalling that $h(\cdot)$ denotes the logistic distribution function.

We turn now to the estimation of UCBs for $\pi_i = \bar{y}_i$ under the three small area models. In **FHtr**, these intervals are standardly based on estimates of mean-squared error (on the transformed scale), i.e., of $E\{\arcsin[\hat{\bar{y}}_i^{\rm BP}]^{1/2} - \arcsin(\sqrt{\bar{y}_i})\}^2$. Two such estimates are respectively a crude one $mse_i^o = \hat{\gamma}_i/(4\,n_i^*)$ and a 'higher-order correct' one (Datta and Lahiri 2000, Rao 2003) defined as

$$mse_i \;=\; mse_i^o + (1-\hat{\gamma}_i)^2\,\mathbf{x}_i^{tr}\,\hat{\Sigma}_\beta\,\mathbf{x}_i + 2(1-\hat{\gamma}_i)^2\,\hat{\Sigma}_{\sigma_u^2}\Big(\frac{\arcsin(\sqrt{y_i^*/n_i^*}) - \mathbf{x}_i^{tr}\,\hat{\beta}}{\hat{\zeta}_i}\Big)^2 \quad (11)$$

given in terms of the notations

$$\hat{\zeta}_i \;=\; \hat{\sigma}_u^2 + \frac{1}{4n_i^*}\,, \quad \hat{\gamma}_i \;=\; \frac{\hat{\sigma}_u^2}{\hat{\zeta}_i}\,, \quad \hat{\Sigma}_\beta \;=\; \Big[\sum_{i=1}^m \frac{\mathbf{x}_i\mathbf{x}_i^{tr}}{\hat{\zeta}_i^2}\Big]^{-1}, \quad \hat{\Sigma}_{\sigma_u^2} \;=\; 2\Big[\sum_{i=1}^m \hat{\zeta}_i^{-2}\Big]^{-1}$$

Based on approximate large-sample normality of $\arcsin[\hat{\bar{y}}_i^{\rm BP}]^{1/2} - \arcsin(\sqrt{\bar{y}_i})$, an approximate level $1-\alpha$ UCB for $\bar{y}_i$ is

$$\mathbf{FHtr}.\mathbf{UCB}_i \;=\; \sin^2\Big(\hat{\gamma}_i\,\arcsin(\sqrt{y_i^*/n_i^*}) \;+\; (1-\hat{\gamma}_i)\,\mathbf{x}_i^{tr}\,\hat{\beta} \;+\; z_\alpha\,\sqrt{mse_i}\Big) \quad (12)$$

However, there are no formula-based numerical methods to estimate UCBs in the small-area models **GLMM** and **BBIN**. Such estimates are instead obtained by a parametric bootstrap approach which has been found to work well in the Fay-Herriot case.

2.3 Small-Area Bootstrap UCB Estimators

To construct confidence intervals or bounds for a small-area target $\pi_i = \bar{y}_i$ in terms of a small-area estimator $\hat{\bar{y}}_i^{\rm BP}$, all existing methods rely on estimates of quantiles of centered and possibly scaled quantities

$$R_i \;=\; \psi(\hat{\bar{y}}_i^{\rm BP}) \;-\; \psi(\bar{y}_i) \qquad \text{or} \qquad S_i \;=\; R_i/\{\hat{V}(\psi(\bar{y}_i^{\rm BP})\}^{1/2} \quad (13)$$

for a fixed smooth strictly increasing function ψ, where $\hat{V} \equiv \hat{V}(\psi(\bar{y}_i^{\rm BP}))$ is chosen as a consistent estimator of the model-based variance of $\psi(\bar{y}_i^{\rm BP}) - \psi(\bar{y}_i)$ and $\bar{y}_i^{\rm BP}$ denotes the best predictor expressed in each model as a function of *known* parameters. In the present context of small-area models, the transforming function ψ is taken to be $\psi(x) = \arcsin(\sqrt{x})$ for **FHtr** and $\psi(x) = x$ for **GLMM** and **BBIN**.

To obtain confidence intervals for $\pi_i = \bar{y}_i$ (sometimes distinguished by the term *prediction intervals* when the quantities π_i are random), we need to find approximate quantiles for the distribution of R_i or S_i. Under conditions (which hold under our three models) guaranteeing uniformly over neighborhoods of parameter values that S_i is approximately distributed as standard normal, the distributions of R_i and S_i are approximately the same as if the data were generated afresh with parameter values fixed at the MLEs $\hat{\vartheta}$ from the observed data. Thus we generate Monte Carlo or *parametric bootstrap* replicate data-vectors $\mathbf{y}^{*(b)} \equiv \{y_i^{*(b)}\}_{i=1}^m$ independently for $b = 1, \ldots, B$ from the same model (**FHtr** or **GLMM** or **BBIN**) assumed to generate $\{y_i^*\}_{i=1}^m$, but with parameters fixed at the MLEs $\hat{\vartheta}$ ($\hat{\beta}$ together with $\hat{\sigma}_u^2$ or $\hat{\sigma}_v^2$ or $\hat{\tau}$), and denote by $R_i^{*(b)}$ and $S_i^{*(b)}$ the random variables defined from $\{y_i^{*(b)}\}_{i=1}^m$ exactly as R_i and S_i were defined from $\{y_i^*\}_{i=1}^m$.

Then if m and B are large, it follows from standard bootstrap theorems (Shao and Tu 1995, Sec. 3.2 and 6.4.4; Hall and Maiti 2006 Sec. 4) that the α_1 and

$1-\alpha_2$ quantiles q_{i1}^R, q_{i2}^R of R_i and q_{i1}^S, q_{i2}^S of S_i (for fixed ϑ) are consistently estimated by the corresponding empirical quantiles q_{ij}^{*R}, q_{ij}^{*S} obtained from the $[B\alpha_1]$ and $[B(1-\alpha_2)]$ order statistics of the respective samples $\{R_i^{*(b)}\}_{b=1}^B$ or $\{S_i^{*(b)}\}_{b=1}^B$. 'Consistency' in this context means that the ratios of the true and estimated quantiles converge to 1.

Then the asymptotically level $1-\alpha_1-\alpha_2$ confidence statement

$$q_{i1}^R \le \Big[\psi(\hat{\bar{y}}_i^{\rm BP}) - \psi(\bar{y}_i)\Big] \le q_{i2}^R$$

translates into the level $1-\alpha_1-\alpha_2$ confidence or prediction interval

$$\bar{y}_i \in \Big[\psi^{-1}\Big(\psi(\hat{\bar{y}}_i^{\rm BP}) - q_{i2}^R\Big),\ \psi^{-1}\Big(\psi(\hat{\bar{y}}_i^{\rm BP}) - q_{i1}^R\Big)\Big]. \tag{14}$$

Similarly, the asymptotically level $1-\alpha_1-\alpha_2$ confidence statement $q_{i1}^S \le S_i = R_i/\hat{V}^{1/2} \le q_{i2}^S$ translates into the confidence or prediction interval

$$\bar{y}_i \in \Big[\psi^{-1}\Big(\psi(\hat{\bar{y}}_i^{\rm BP}) - \hat{q}_{2i}^S\,\hat{V}^{1/2}\Big),\ \psi^{-1}\Big(\psi(\hat{\bar{y}}_i^{\rm BP}) - \hat{q}_{1i}^S\,\hat{V}^{1/2}\Big)\Big] \tag{15}$$

The recipe we have described here to obtain bootstrap-based quantiles and confidence intervals is the one we follow in calculating UCBs for the three models **FHtr**, **GLMM**, and **BBIN**, with respective choices of $\psi(x) = \arcsin(\sqrt{x})$, x, and x, and with $\alpha_1 = \alpha$, $\alpha_2 = 0$. In the **FHtr** case (after taking account of the transformation of scale), this is exactly the bootstrap algorithm proved by Chatterjee et al. (2008) to yield confidence or prediction intervals with coverage probabilities accurate to an order of $m^{-3/2}$. This is an accuracy greater than can be claimed for the UCB (12) based on mse_i estimators of mean-squared prediction error (on transformed scale). For the **GLMM** and **BBIN**, the precise results of Chatterjee et al. (2008) are not available, but the parametric bootstrap idea still provides (as argued in general terms by Hall and Maiti 2006) coverage accuracy of order $1/m$ for the bootstrap-based UCBs. Parametric bootstrap intervals for generalized linear models like **GLMM** and **BBIN** are given by Hall and Maiti (2006), but as remarked by Chatterjee et al. (2008, Remark 8 in Sec. 3), the intervals provided there are prediction intervals not for the target $\bar{y}_i$ based on the statistic $\hat{\bar{y}}_i^{\rm BP} - \bar{y}_i$, but rather for $\bar{y}_i$ based on $\mathbf{x}_i^{tr}\,\hat{\beta} - \bar{y}_i$.

3. Effective Sample Size

Determining the method is not the only modeling priority. Another is to determine the effective sample size n_{gi}^* to use in the modeling, where g is the index for the higher-level aggregate (group of areas) and i is the index for individual areas included in g. (Throughout this section, we use the term n_{gi}^* to denote exactly the same quantity elsewhere denoted n_i^*, in order to emphasize the grouping of areas within a higher-level aggregate.) The effective sample size is a measure of how many 'independent' units of information contribute to the estimate. In complex samples, sampled units in an area may be correlated, thus reducing the amount of independent information. The proposed methods below are intended to reduce the sample size for an area to more accurately reflect the equivalent amount of independent information. These adjustments may greatly affect the power for hypothesis testing and variance estimation.

3.1 Ratio-Adjusted Sample Size

The Ratio-Adjusted Sample Size (RA) is the simplest scheme for reweighting sample cases in each area. It involves taking the weights that are already present in the small area, and scaling them to a fixed known total so that the sum of the adjusted weights equals the unweighted sample size (over the whole sample). The relative ratio between sample weights remains the same. Then the effective sample size, n^*_{gi}, for an area is the sum of the adjusted sample weights:

$$n^*_{gi} = \sum_j w^*_{gij} = \frac{n}{\hat{N}} \sum_j w_{gij} \tag{16}$$

where $\hat{N} = \sum_{gij} w_{gij}$ estimates the overall population size by the sum of all the sampling weights, n is the total number of records in the dataset, and w_{gij} is the weight for household j within area i which in turn lies within aggregate group g.

This adjustment essentially represents a synthetic estimate of sample size, with no modeling involved. The sample size is adjusted down in (16) so that the predictive power of models can be assessed as though the sample were unweighted. However, this method may not accurately capture the variance of the sampling scheme. An alternative is to use a *design effect*, as in the two options below.

3.2 Constant Design Effect

A sample size divided by a design effect for a particular attribute yields the equivalent sample size that would produce the same design-based variance estimates for the attribute under the assumption of simple random sampling. When reliable area-level design effects are not available, one may use a *constant design effect* assumption, that is, one may start with a higher-level design effect (for a group of areas) and assume that they are all equal to the area-level design effect. For example, design effects for states can be calculated and then assumed to be the same as the design effect of all counties (or places) within the state. First let $n_g = \sum_i n_{gi}$ be the sum of sample-sizes over all areas i within group g, and let $\hat{p}_g$ be the estimated group rate. Then we use the design effect to adjust unweighted sample size n_{gi} for area i, by obtaining an aggregate-level variance estimator $\hat{V}_g$ and using it to find an effective sample size n^*_{gi}:

$$\begin{aligned} \hat{V}_g &= \mathrm{DEFF}_g \ \hat{p}_g(1-\hat{p}_g) \,/\, n_g \\ \mathrm{DEFF}_g &= n_g \hat{V}_g / \{\hat{p}_g\,(1-\hat{p}_g)\} \\ n^*_{gi} &= n_{gi} \,/\, \mathrm{DEFF}_g \end{aligned}$$

3.3 Weight-Adjusted Design Effect

The weight-adjusted design effect is similar to the constant DEFF, but it is able to take into account the variability of the weights within the area. This method was derived from a model-based justification of Kish's formula for design effects in clustered data (Gabler et al, 1999). The weight-adjusted design effect calculates the design effect, b_g, after taking unequal sampling weights into account. The use of such a design effect is also motivated by Hawala and Lahiri (2010). We find the weight-adjusted effective sample size n^*_{gi} through the aggregate-level variance estimator $\hat{V}_g$ by:

$$\begin{aligned}
\hat{V}_g &= b_g\, c_g\, \hat{p}_g\,(1-\hat{p}_g) \qquad \text{where} \qquad c_g = \Sigma_{i,j}\, w_{gij}^2/(\Sigma_{i,j}\, w_{gij})^2 \\
b_g &= \hat{V}_g/\{c_g\, \hat{p}_g\,(1-\hat{p}_g)\} \\
n_{gi}^* &= \frac{1}{b_g\, c_{gi}} = \frac{c_g}{c_{gi}}\frac{n_g}{\mathrm{DEFF}_g} \qquad \text{where} \qquad c_{gi} = \Sigma_j\, w_{gij}^2 \,/\, (\Sigma_j\, w_{gij})^2
\end{aligned}$$

Recall that i indexes counties, j indexes HUs, and summations over i range over all i within group g. As with the constant design effect, the weight-adjusted design effect is restricted by the design assumption of the adjustment factor carrying down. However, it does take into account weight variability in ways that the other methods do not.

4. Census Coverage Measurement Application

The Census Coverage Measurement (CCM) program is an effort to measure decennial census accuracy by comparison with an independent population survey, also called CCM. The CCM produces Correct Enumeration (CE) and Erroneous Enumeration (EE) rates – the proportion of Census records that are correct and incorrect, respectively – tabulated across different geographic boundaries. The CCM studies these rates for both the individual and housing-unit (HU) records; this study focuses on HU EE-rate estimation.

One of the goals of CCM is to estimate the components of enumeration among HUs: totals and rates of CEs and EEs for different areas, broken down into smaller categories. The Census Bureau publishes those estimates, along with the associated standard errors, for counties and places of over 500,000 people. These estimates are known as the CCM Housing Unit production estimates, and the specific jurisdictions are known as production counties or production places.

The 2010 CCM used a probability-based sample of over 170,000 Housing Units (HUs) across the United States and Puerto Rico. It features 1,728 counties and 2,630 places, of which 128 counties and 33 places are in production. The quantities observed in the CCM are survey-weighted ratio estimators $\hat{\bar{y}}_i$ as in (1).

The rates for Erroneous Enumeration are typically small. The 2010 overall EE rate is 2.7% (Keller and Fox 2012) and there were fifteen production counties and four production places that had an estimated rate $\hat{\bar{y}}_i < .001$, which CCM considered to be zero.

The goal of this project was to produce reasonable estimates of uncertainty for those nineteen areas. However, more broadly there was interest in developing a paradigm for handling similar situations in the future, as this problem applies across different operations at the Bureau.

5. Component Modeling with CCM

The first step was to develop county and place models for Erroneous Enumerations, but this paper only addresses county-level results. EE rates, household weights, and many potential covariates were extracted from the CCM data files. For this project, an enumeration is labeled correct if it is classified as type ACPRHUCE, PRHUSB, or PRHUGE (see Table 1) and it is labeled as erroneous if classified as type APRHUDUP or PRHUOTH.

5.1 Initial Modeling and Testing

The models were fit at the county and place levels simultaneously. All areas in sample were used for building and estimating them. Because CCM uses a sample of block groups, the candidate variables were taken from the 2010 decennial census, which included demographic and census operational variables for each area.

We evaluated many fixed-effects candidate models using likelihood testing and the Bayesian Information Criterion (BIC), $-2 \cdot \ln(L) + k \cdot \ln(n)$, where L is the likelihood, k is the number of parameters, and n is the total records in sample. This formula assesses a penalty of $\ln(n)$ for each extra parameter. Model checking methods identified the same model for both counties and places: a logistic regression model, using an effective sample size defined by the ratio-adjusted scheme of Section 3.1. The model included these five covariates:

- The logistic-transformed rate of *correct enumerations for the state*, as a synthetic estimator;
- The arcsin-square-root of the rate of *single-unit households* for the area;
- The arcsin-square-root of the rate of *multi-unit households* for the area;
- The arcsin-square-root of the rate of *urban households* for the area;
- The arcsin-square-root of the *enumeration rate* for the area.

The model chosen had the lowest value of BIC for both counties and places. Subsequent tests to assess the quantile fit of models subjectively confirmed our choice of best model.

The range of various weight parameters for counties is given in Table 2. The effective sample sizes show a wide range that depends on the method used. The range of n^* using adjusted weight is very different from the design effect schemes (which are similar to each other).

The sum of the effective sample size for all areas is 46,690 for the constant design effect method and 43,066 for the weight-adjusted design effect method. These sums are both approximately a quarter of the sum under the allocated weight scheme (which is fixed at the CCM sample of 172,503).

6. Results

The results vary by UCB estimation method and the type of effective sample size chosen, but we present results only for the weight-adjusted effective sample size described in Section 3.3. We determined that the weight-adjusted effective sample size did a good job of capturing cross-county variations in HU clustering. These

Table 1: CCM Component Variables

Variable Name	Definition
ACPRHUCE	CE in block cluster
PRHUSB	CE in nearby blocks
PRHUGE	CE with geocoding error
APRHUDUP	EE due to duplication
PRHUOTH	EE due to other

sample sizes are usually a little bit larger than the constant design effect sample sizes, with a maximum difference of about 20%. However, the differences in UCBs due to the choice of Constant design effect versus Adjusted-Weight design effect are typically no more than about .01. By contrast, the ratio-adjusted effective sample sizes were much larger, leading to much smaller one-sided confidence intervals which were too narrow to be trustworthy.

Viewed over the Production counties, i.e., those with total population $\geq 500,000$, the Cell-based UCBs tend to be larger than the UCBs from small-area models, with the **GLMM** model-based bounds generally lying above the **FHtr** bounds, which are close to but tend to be slightly larger than the **BBIN** bounds. This pattern can be seen in Figure 1. (The UCBs produced by the bootstrap methods for the small-area models were based on $B = 2500$ bootstrap replications. Generally, at least 1000 replicates are needed to provide the desired level of accuracy).

To get a clearer idea of the comparison between the UCBs on individual large counties, the scatterplot in Figure 2 shows that UCB size is ordered **GLMM** $\succ$ **FHtr** $\succ$ **BBIN**, where '$\succ$' means 'generally greater than'.

One of the questions which motivated our study was the appropriate method to provide upper confidence bounds on direct estimated (HU EE) rates of 0. Among the 15 production counties whose direct EE rates were 0, we see in Table 3 a somewhat different relationship among the UCBs provided by the four methods. Now the Cell-based method provides the narrowest intervals for $n_i^* \geq 30$. This makes some sense because the Cell-based method does not borrow strength from counties with larger point estimates, as the others do, and therefore its upper bound will be smaller when the estimated rate is. However, we caution the reader that the anomalously small Cell-based bounds for the counties with estimates close to 0 are to some extent an artifact of having used the anticonservative arcsin square root method; when the method is modified as suggested below (3), the Cell-based UCBs are no longer smaller than the GLMM-based UCBs for effective sample sizes less than 118.

This paper has been concerned with UCBs, which are understood to bracket the direct estimators of survey proportions, but one may also ask about the point estimators generated by small-area methods. These are displayed in boxplot format in Figure 3. The point estimates for the different methods reflect similar means across all methods but narrower ranges for the model-based ones, while the UCBs from Figure 1 reflect lower means for the model-based methods. We propose to use these estimators in conjunction with prediction intervals based on $\mathbf{x}_i^{tr}\hat{\beta}$ for model

Table 2: Mean and Quartiles for Effective n and Associated Ratios based on counties with ≥ 20 HUs in sample.

	Mean	1st Q	Median	3rd Q
$n^{\wedge}$	119.4	26.4	52.4	115.4
$n^{\dagger}$	30.3	6.4	12.9	28.2
$n^{\ddagger}$	32.8	7.5	14.3	30.2
$n^{\ddagger}/n^{\dagger}$	1.2	1.1	1.1	1.2

$n^{\wedge}$ = Ratio-Adjusted sample size, $n^{\dagger}$ = Constant DEFF sample size, $n^{\ddagger}$ = Weight-Adjusted DEFF sample size.

Table 3: UCBs of HU EE Rates for Larger Counties where $\hat{\hat{y}}_i < 0.001$.

Cty n_i^*	FH.pred	Cell	FHtr	GLMM	BBin
10.4	.0105	.0638	.0240	.0339	.0183
11.7	.0108	.0569	.0233	.0335	.0188
20.1	.0168	.0332	.0320	.0431	.0233
30.8	.0123	.0218	.0265	.0415	.0228
60.7	.0064	.0111	.0163	.0272	.0144
60.8	.0021	.0111	.0087	.0193	.0103
98.9	.0061	.0068	.0151	.0276	.0145
101.9	.0027	.0066	.0098	.0217	.0110
118.8	.0073	.0057	.0165	.0335	.0169
120.3	.0013	.0056	.0067	.0154	.0079
135.9	.0052	.0101	.0132	.0245	.0124
164.5	.0022	.0041	.0080	.0192	.0091
171.0	.0031	.0040	.0092	.0234	.0108
186.7	.0014	.0067	.0060	.0153	.0073
188.8	.0036	.0036	.0096	.0230	.0106

checking, as Slud (2012) did with the **FHtr** model in another data application.

7. Conclusion and Future Research

Because the UCB estimates generated from all methods were roughly similar, CCM staff made the decision to use the Cell-based method in releasing the production county results in May 2012. It is the most easily explained method and does not rely on a regression model. The published estimates reflect the upper confidence bounds from the Cell-based method. CCM does not publish bounds, but did publish pseudo-standard errors that reflected the same one-sided 95 percent confidence bound when applied on the probability scale.

The methods presented here suggest that model-based UCBs estimated by small-area models using a parametric bootstrap approach can be used effectively to bracket the unknown area-level proportions corresponding to small estimated survey proportions. Before such UCBs could be released in official data products, additional research assessing model fit must be undertaken. Such research should probably also include some sensitivity checking of the results to the choice of the method of defining effective sample sizes.

7.1 Future Plans

The general problem discussed here is applicable to other work done at the Census Bureau. Ideally, this research can provide a framework to others working on similar tasks, perhaps through the development of an R package that would incorporate these methods.

The need for model checking in specific applications of these methods has already been emphasized above. In addition, specific topics for further methodological research which would enhance the applicability and value of the methods discussed here include:

(i) alternative numerical procedures for calculating and approximating bootstrap quantiles, including a method for justifying the choice of a number B of bootstrap iterations;

(ii) investigating which features of specific small-area models tend to lead to larger or smaller estimated UCBs;

(iii) exploring alternative choices of effective sample-size definitions, and what might justify them in specific survey applications.

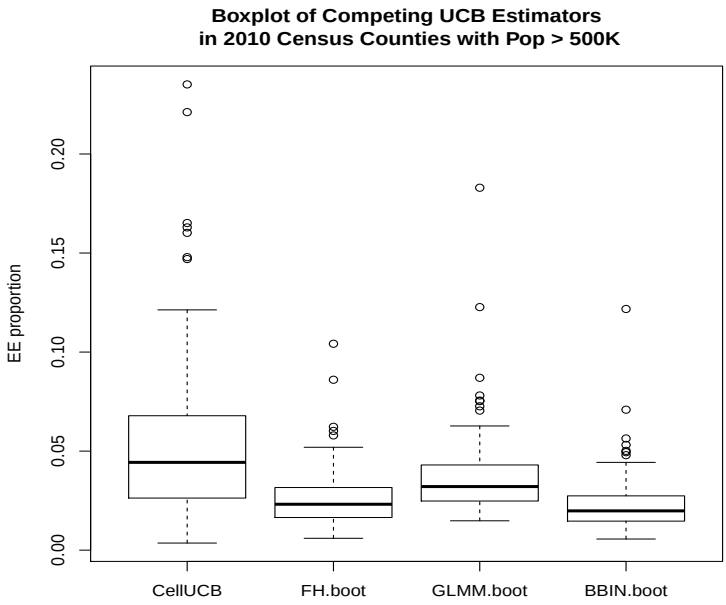

Figure 1: Boxplot of UCBs produced on 128 counties with total population $\geq$ 500,000, by the Cell-based and 3 Small Area methods, using the bootstrap approach described in Section 2.3.

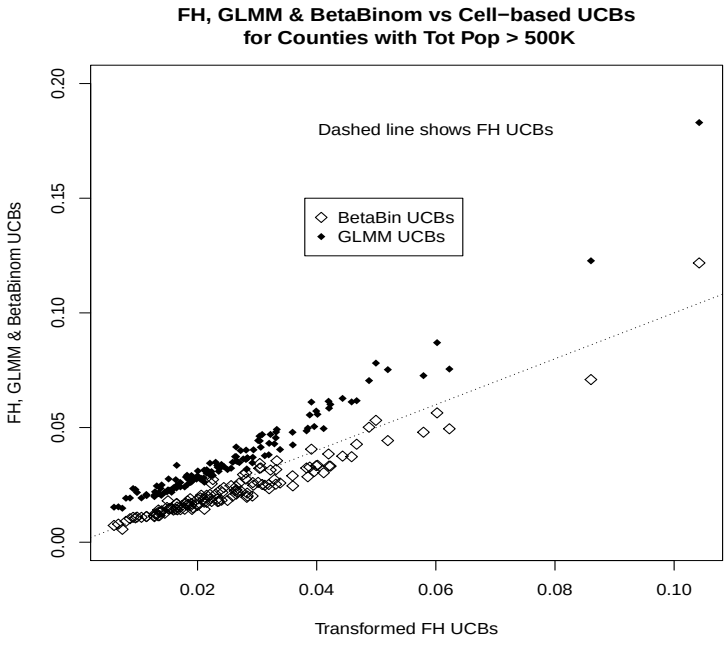

Figure 2: Scatterplot of UCBs produced on 128 counties with total population $\geq$ 500,000, by the **GLMM** and **BBIN** methods, each plotted against those produced by the **FHtr** method. All of these UCBs were estimated by the bootstrap approach described in Section 2.3.

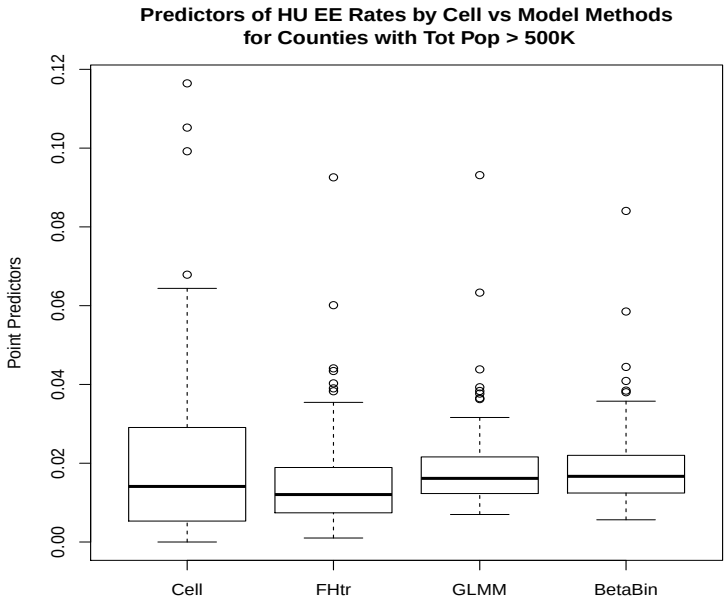

Figure 3: Boxplot of predictors for 128 counties with total population $\geq$ 500,000.

References

Chatterjee, S., Lahiri, P. and Li, H. (2008), *Parametric Bootstrap Approximation to the Distribution of EBLUP and Related Prediction Intervals in Small-area Estimation.* Annals of Statistics, 36, 1221-1245.

Datta, G., and Lahiri, P. (2000), *A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems.* Statistica Sinica, 10(2000), 613-627.

Fay, R. and Herriot, R. (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.

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.

Hall, P. and Maiti, T. (2006), *On Parametric Bootstrap Methods for Small Area Prediction.* Journal of the Royal Statistical Society: Series B, 68, 221-238.

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 Survey Research Methods Section, American Statistical Association.

Jiang, J. and Lahiri, P. (2006), *Mixed Model Prediction and Small-area Estimation (with Discussion).* Test 15, 1-96.

Keller, A. and Fox, T. (2012), *2010 Census Coverage Measurement Report: Components of Census Coverage for Housing Units in the United States.* DSSD 2010 Census Coverage Measurement Series #2010-G-06.

Liu, Y. and Kott, P. (2009), *Evaluating Alternative One-Sided Coverage Intervals for a Proportion.* Journal of Official Statistics, 25, 569-588.

Prentice, R. (1986), *Binary Regression using an Extended Beta-binomial Distribution, with Discussion of Correlation Induced by Covariate Measurement Errors.* Journal of the American Statistical Association, 81, 321-327.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rao, J.N.K. (2003), **Small Area Estimation**. Wiley.

Shao, J. and Tu, D. (1995), **The Jackknife and Bootstrap**. Springer.

Slud, E. (2012), *Assessment of Zeroes in Survey-Estimated Tables via Small-Area Confidence Bounds.* Journal of the Indian Society of Agricultural Statistics, 66, 157-169.