An Evaluation of Synthetic Small Area Census Coverage Error Using a Random Effects Model[1]

Donald Malec and Jerry Maples
U.S. Bureau of the Census
Washington D.C. 20233

Key Words: Dual System Estimation, Capture/Recapture, Bayesian Inference

1. Introduction

The Accuracy and Coverage Evaluation Survey (A.C.E.) covers many local areas such as states, congressional districts and census field offices. However, there is often not a large enough sample to provide precise design-based estimates of local coverage. As an alternative to direct estimates, the synthetic method of small area coverage estimation applies national post-stratum level coverage estimates to local census counts, instead of using local coverage estimates. Although it is acknowledged that the coverage rate will vary geographically within post-strata, a coverage estimate based on a synthetic estimate that ignores this level of geography is still perceived as being better than reporting no coverage estimate at all.

An alternative to using either a synthetic estimate or a design-based estimate is to use a borrowing strength estimator, i.e., one that combines features of both a design-based estimator and a synthetic estimator. Borrowing strength estimators have been used extensively in small area estimation (see, e.g. Rao, 2003). Although the borrowing strength methods proposed will require intensive computing, standard algorithms for doing so are available. We investigate the use of borrowing strength estimators that are appropriate to the coverage estimation problem by incorporating a random effect model to help assess small area variability and aid in prediction.

We choose the local census offices (LCOs) as the small areas to be initially investigated. There are 540 LCOs covering the U.S. which can provide information on the geographic variation and patterns across the U.S. Being administrative units for nonresponse followup data collection, information about LCO coverage could be of use for planning purposes. Estimates at the LCO level (by post-strata) can still be synthetically carried down to smaller (or different) areas and can also incorporate any LCO-specific covariates.

The nature of small area estimation precludes comparison of estimates against the truth because only small sample sizes are available for making design-based estimates. Our primary evaluation will be to determine whether our model can describe the A.C.E. data better than a comparable synthetic-estimation model and whether the incorporation of the random effects results in substantially different estimates of coverage. We will evaluate a key assumption of the synthetic estimation model that coverage rates (i.e., correct enumeration rates and matching rates) are relatively constant across small areas (LCOs). The first part of the evaluation will be accomplished by nesting a synthetic estimation model in our random effects model and then assessing whether the variance components are negligible, or not (based on posterior confidence intervals). If the variance components are negligible, then this would be evidence that the synthetic assumption is valid. The second part will be accomplished by making estimates from the two models and assessing their differences. Ultimately, we plan on providing evidence whether, or not, the addition of small area random effects could improve synthetic estimation.

This project is exploratory in nature. We do not evaluate the fit of the data to the model. We simply have embedded a synthetic estimation model into a larger one in order to evaluate possible improvements in using a larger, random effects, model. The main contribution of this project is the evaluation of the synthetic model, which has a large number of parameters, and implementation of an estimation procedure which takes into account that the sample selected is not from a simple random sample. Model fitting and careful evaluation of the sample design adjustment will be pursued only if this project indicates that this method may offer gains over synthetic estimation.

[1]This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The 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 authors would like to thank colleagues at the Bureau, especially Tom Mule and Eric Schindler for their help on the project and additionally thank Stephen Ash, Donna Kostanich and Smanchai Sae Ung for reviewing the manuscript.

2. Census Coverage Estimation Overview: From the National Level to Small Areas

Conceptually, coverage estimation follows a capture/recapture approach (see, e.g. Seber(1987)) That is, the first capture, of N_1 people, is from the census and the second capture, of N_2 is from the coverage survey. Given that the chance of being captured is independent between sources and that M persons match, an estimate of the total population is: $N_1 N_2/M$, providing a coverage estimate of N_2/M. There are other sources of error present in Census coverage that require making changes to the preceeding estimate, (see, e.g. Hogan (2003, 1993)). Four primary sources of error are: 1) the occurrence of erroneous enumerations in census, 2) lack of information to match records, 3) availability of coverage survey data for only a sample of areas (due to cost limitations) and 4) heterogeneous probability of being selected in sample. Each Census record does not match a single person due to the fact that there are fictitious persons and duplicates in the Census, hence N_1 is not known. In addition, some census records are believed to represent real people but do not contain enough (extraneous) information to determine whether, or not, they match to anyone in the coverage sample. Since N_1 cannot be determined, an estimate of the total number of persons with census records that are matchable, is used instead. The determination of the correct, matchable enumerations are carried out on only a sample basis, as are the number of persons in the coverage survey, N_2, and the number of matches, M. Taking the first three types of errors into account, the form of the modified capture/recapture estimate looks more like the following:

$$Cen \times P_{DD} \times \hat{P}_{CE}/\hat{P}_M$$

representing the Census count, Cen, the proportion of census records that are matchable, P_{DD}, (the "DD" stands for "Census Data Defined", which refers to census records that do contain enough information on them to warrant any attempt at matching, e.g. imputed records), the sample estimates of the correct enumerations rate in the Census, $\hat{P}_{CE}$, and the sample estimate of the match rate between the data defined records and the coverage survey, $\hat{P}_M$. These estimators are generally referred to as "dual system estimators". The resulting coverage estimator looks like:

$$P_{DD}\hat{P}_{CE}/\hat{P}_M$$

The fourth source of error, heterogeneous capture probabilities, is common in many capture/recapture settings. The bias caused by ignoring this fact is reduced by stratifying the samples in more homogeneous post-strata. For census coverage, the census records (collected in the "e-sample") and the coverage records (collected in the "p-sample") are stratified differently in the U.S. Census Bureau's estimates referred to as "revision II". A final estimate of coverage at the national level is based on producing estimates of population for each post-stratum type, i,j, and then aggregating. An estimate of total coverage would take the following form

$$\frac{\sum_{i,j} Cen_{ij} P_{DDij} \hat{P}_{CEi}/\hat{P}_{Mj}}{\sum_{i,j} Cen_{ij}}.$$

Note that it is assumed the correct enumeration probabilities, P_{CEi}, are homogeneous within the e-sample post-strata, i, and the coverage probabilities, P_{Mj}, are homogeneous within the p-sample post-strata, j. There are 480 post-strata defined for the P-sample. These post-strata are constructed from a cross-classification of the groups within the categories: Race/Hispanic Origin, Tenure, Size of Metropolitan Statistical Area, Type of Census Enumeration Area, Return Rate Indicator, Region, Age,and sex. The e-sample post-strata are not based on the same set of variables. In fact, some sample post-strata cannot be defined for p-sample cases because they are based on census information. The full e-sample consists of 525 post-strata defined by cross classifications of groups based on the following characteristics: Proxy Status, Race/Hispanic Origin Domain, tenure, Household Relationship, Household Size, Type of Census Returns (mail back vs non-mail back), Date of Return (early vs. late), age and sex. The report "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology" (2004) provides the explicit definition of both the e-sample and p-sample post-strata. There are a number of other errors that are accounted for in the coverage error estimation of the Census and this report provides further details on that as well as details on the overall estimation process.

If the coverage is homogeneous within all crossed post-strata, the estimated coverage rates apply to any aggregate level of sample within post-strata. For an arbitrary area, k, one partitions the area by post-stratum and produces a separate estimate resulting in the synthetic estimate of coverage for the area, k,:

$$\hat{CCF}_k = \frac{\sum_{i,j} Cen_{ijk} P_{DDij} \hat{P}_{CEi}/\hat{P}_{Mj}}{\sum_{i,j} Cen_{ijk}}$$

Where Cen_{ijk} is the number of census records in post-strata combination i, j in area k. Note that a synthetic assumption was also applied to the data-defined rate for production estimates.

The explicit aim of this paper is to evaluate this homogeneity assumption, by allowing heterogeneity between LCOs , within post-strata. Before the inclusion of random effects to account for this small area heterogeneity is evaluated, a corresponding model is needed. The following sections describe the model to be used and cover the problems of model fitting using data disproportionately sampled from its population.

3. The Finite Population and Sample Design

Only the aspects of the sample selection that are relevant to this problem will be described. A detailed account of the design can be found in "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology" (2004).

The basic sampling unit is a block cluster which, generally, consists of a typical city block or a collection of geographically contiguous housing units about the same size as a city block. The median target block cluster size is about 30 housing units but ranges considerably between zero housing units and 80+. The population of block clusters totaled over 3.5 million. These block clusters were stratified by State and crossed by whether or not they are located on an American Indian Reservation and, if not, by three levels of estimated size (0-2, 3-79 or 80+ housing units). A sample of about 29,000 block clusters was selected from these strata. Subsequent to this initial selection, oversampling was employed to achieve a variety of objectives such as requiring adequate sample size for each state and for race/ethnicity groups. Double sampling was then employed in most of these primary strata reducing the original sample down to about 11,000 block clusters. For medium and large block clusters, this second phase is a sub-sample from sub-strata constructed from information collected in the original sample on such items as actual measured size of the block clusters (in terms of housing units), matching rates of housing units between the census list of housing units and the independent list used to take the sample, minority percent based on the original 1990 census estimates, as well as block clusters whose actual size were very different from their estimated size. As in the first-phase sample, oversampling is employed. In this second phase, block clusters that had a poor address match rate between the Census address list and the independent sample listing of the follow-up survey were oversampled, as were minority block clusters. Unusually large block clusters were further sampled to reduce the work load. The primary features of the design that are different from a simple random sample of people in the U.S (or of census records) consist of the clustering of individuals in housing units and contiguous geographic areas, the initial stratification based on strata identifiers that may be related to coverage, and sub sampling based on second-phase strata. Of all these features, subsampling based on the initial match between housing units must be accounted, perhaps most of all, to ensure that estimates from the model will not be biased.

4. Population Model

We do not try to completely specify a unit-level population model for the coverage process. Doing so would entail including one capture probability denoting the chance that a person is included in the census frame and another denoting the chance that a person is included in the coverage survey frame, a probability denoting the chance of a erroneous census record, including duplicates and fictitious records, accounting for differential coverage due to geo-coding error, movers and a number of other sources of error in "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology" (2004), report. Instead we specify a relatively simple model for the probability of a match and the probability of a correct enumeration, use the sample design to account for differential selection and correlation and, lastly, condition on the e-sample and p-sample totals to make inference about the coverage as a function of only two types of unknowns, the probability for correct enumeration and the probability of a match.

Model Assumption 1 Define the probability that a person in the sample of matchable census records (e-sample) is a correct enumeration, given they are in small area k, e-sample post-strata, i, as: $p_{(CE,i,k)}$.

Model Assumption 2 Analogously, define the probability that a person in the coverage sample matches a correct, matchable census record, given they are in small area k, p-sample post-strata, j, as: $p_{(M,j,k)}$.

Within LCO, k, the following parameterization of $p_{(CE,i,k)}$ and $p_{(M,j,k)}$ is assumed:

$$p_{(CE,i,k)} = e^{\theta_{CEik}}/(1 + e^{\theta_{CEik}})$$

$$p_{(M,j,k)} = e^{\theta_{Mjk}}/(1+e^{\theta_{Mjk}}) \quad (1)$$

We will require a complete parametric model, in order to obtain confidence intervals based on posterior distributions. A parametric model will be built, within an LCO based on the following reasoning:
1) Starting with the specification of the marginal means in equation (1), a pseudo-likelihood is formed by treating each observation as independent but using the sample weights so that the sampled log-likelihood is a design-unbiased estimate of the population log-likelihood (based on independence).
2) This pseudo-likelihood is scaled down, exponentially, to reflect the effective sample size due to clustering, and increased variability due to some design features not being included in the model.

Note that the use of a pseudo-likelihood is often used in modeling survey data, see Skinner et al. (1989) for more details. The following provides additional details on the justification of this approach.

4.1 Step 1: Pseudo-Likelihood

Fixing the parameters $p_{(CE,i,k)}$ and $p_{(M,j,k)}$, the following pseudo-likelihood can be used to determine their MLEs:

$$\prod_i p_{(CE,i,k)}^{m_{CE,i,k}}(1-p_{(CE,i,k)})^{n_{CE,i,k}-m_{CE,i,k}} \times \prod_j p_{(M,j,k)}^{m_{M,j,k}}(1-p_{(M,j,k)})^{n_{M,j,k}-m_{M,j,k}} \quad (2)$$

where, $m_{CE,i,k}$ and $n_{CE,i,k}$ are, possibly rescaled, unbiased estimators of the total number of correct enumerations and population total, respectively, in e-sample post-strata i and LCO k. The terms $m_{M,i,k}$ and $n_{M,i,k}$ are defined analogously for the p-sample. These estimates incorporate missing value adjustments as well as the Revision II adjustments.

As an illustration for the e-sample term, $m_{CE,i,k}$ and $n_{CE,i,k}$ are unbiased estimates of the total number of correct enumerations rescaled by $a_{CE,i,k}$:

$$m_{CE,i,k} = \sum_{b\in s} a_{CE,i,k}\; w_b\; m_{CE,i,k,b}$$

and

$$n_{CE,i,k} = \sum_{b\in s} a_{CE,i,k}\; w_b\; n_{CE,i,k,b},$$

where, b denotes a block/cluster sampling unit, $b \in s$ denotes all block clusters in sample, w_b is the sampling weight, $m_{CE,i,k,b}$ and $n_{CE,i,k,b}$ are the actual counts of correct enumerations and counts of census records observed in block cluster, b. Lastly, $a_{CE,i,k}$ is any arbitrary positive constant which will be specified later in order to force the variances from the MLEs to be comparable with those resulting from a cluster sample.

Note that the log-pseudo likelihood is an unbiased estimator of a population log-likelihood based on an independent sample of people in the p-sampled and an independent sample of census records in the e-sample. This assumption clearly may not represent the actual random structure of the sample, as persons (or census records) are likely to be correlated within housing units and within blocks. Following Huber(1967), the MLEs from the above pseudo-likelihood will still be consistent estimates of θ_{CEik} and θ_{Mjk} as long as the total expectation of the pseudo-log likelihood, under the correct population model is maximized at θ_{CEik} and θ_{Mjk}, which is true in this case.

4.2 Step 2: Scaled Pseudo-Likelihood

Following the pseudo-likelihood approach, we pick a constant $a_{CE,i,k}$ to adjust the sample size so that variances of MLEs based on the pseudo-likelihood match up to variances based on the sample design/model specification. As pointed out by Graubard and Korn (2002), model based estimation from sampled data must account for variability of the design parameters, if they are not part of the model. We obtain variance estimates of MLE estimators by boot-strapping block clusters, including bootstrapping the initial strata and substrata identity. The use of the bootstrap should account for variability of the sample selection, variability of the assignment of block clusters to the design parameters and model-based correlation within block-clusters. In this present work, a conservative estimate of variance has been used. In particular, the average variance of rates determined within LCO by collapsed post-stratum have been used. (See Appendix A for collapsed post-strata definitions). This approach is conservative because it includes the variability due to changes of post-strata sample size within collapsed post-strata. We have also assumed independence between block-clusters.

Note that this approach is taken because the task of completely modeling the individual responses in lieu of correlation within clusters, and the effect of the sample design is daunting. Also, assuming asymptotic Normality for each component is not appropriate since the sample sizes in many cases are too small. Although there are 521 LCOs out of 540 that have some data, data is needed at the LCO by e-sample and p-sample post-strata level in order to account for both the original post-strata effects and LCO effects. There are 125,212 of these e-sample

cells and 46,059 of these p-sample cells with non-zero sample sizes. These e-sample cells have sample sizes that range between 1 and 610, 25% have only a sample of one, 50% have a sample of 3 or less and 75% have a sample size of 6 or less. These p-sample cells have sample sizes that range between 1 and 647, 25% have only a sample of 2, 50% have a sample of 6 or less and 75% have a sample size of 16 or less.

5. Random Effect Model

Many, if not all, of the parameters, θ_{CEik} and θ_{Mik} will not have enough sample size to provide estimates to assess the variability across LCOs within post-strata. In order to assess the small area variation, random effects models are used. Within an LCO, k, by e-sample post-strata cell, i, and p-sample post-strata cell, j, we specify the following model as:

$$\begin{aligned} \theta_{CEik} &= \beta_{CEi} + \mu_{CEk} + \alpha_{CEik} \\ \theta_{Mjk} &= \beta_{Mj} + \mu_{Mk} + \alpha_{Mjk} \end{aligned} \tag{3}$$

The random effects are assumed to be normally distributed, with unknown variances that will be estimated from the data. Specifically, defining $(\mu_{CEk}, \mu_{Mk})=\underline{\mu}_k$, the LCO random effects are given the distribution:

$$\mu_k \sim N(\underline{0}, \Sigma) \tag{4}$$

The terms, α_{CEik} and α_{Mjk} are added to represent model error that may be differentially present within post-strata cells. Although adding these terms will not account for all modeling error (modeling at the individual level would be needed), these terms are added as a way to assess possible future work where more modeling may be needed. Specifically, the model errors are given the distribution:

$$\begin{aligned} \alpha_{CEik} &\sim N(\underline{0}, \gamma^2_{c_e(i)}) \\ \alpha_{Mjk} &\sim N(\underline{0}, \gamma^2_{c_p(j)}) \end{aligned} \tag{5}$$

where $c_e(i)$ collapses the original 525 e-sample post-strata into 11 cells and $c_p(j)$ collapses the original 480 p-sample cells into 8, (specified in Appendix A). As summarized in section 6., there is usually very little sample available to model variance components at the LCO by post-strata level. The original 525 e-sample post-strata and the original 480 p-sample post-strata are still incorporated into this model as fixed effects, in order to evaluate the synthetic model.

If a model were used to produce the Revision II estimates, it could be based on the assumption that $\theta_{CEik}=\beta_{CEi}$ and that $\theta_{Mj}=\beta_{Mjk}$, i.e. post-strata are homogeneous across LCOs.

6. Estimation and Inference

We use a Bayesian approach with non-informative priors to assess the variance components $\underline{\gamma}^2_{c_e}$, $\underline{\gamma}^2_{c_p}$ and Σ.

The likelihood component includes both the Scaled Pseudo-likelihood of (2)described in sections and and the random effects distributions specified in (4) and (5). These distributions, combined together, are proportional to:

$$\begin{aligned} &\prod_k \Bigg(\prod_i p_{(CE,i,k)}^{m_{CE,i,k}} (1-p_{(CE,i,k)})^{n_{CE,i,k}-m_{CE,i,k}} \\ &\times \prod_j p_{(M,j,k)}^{m_{M,j,k}} (1-p_{(M,j,k)})^{n_{M,j,k}-m_{M,j,k}} \\ &\times |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}\mu_k' \Sigma^{-1} \mu_k} \\ &\times \prod_i \gamma^{-1}_{c_e(i)} e^{-\frac{1}{2}\alpha^2_{CEik}/\gamma^2_{c_e(i)}} \prod_j \gamma^{-1}_{c_p(j)} e^{-\frac{1}{2}\alpha^2_{Mjk}/\gamma^2_{M(j)}} \Bigg) \end{aligned} \tag{6}$$

The aim of small area estimation is to provide estimates that include the effects: $\underline{\mu}_k$, α_{CEik} and α_{Mjk}. We will use MCMC methods to obtain their posterior distribution and, in turn, construct estimated coverage rates based on the posterior distribution of:

$$\hat{CCF}_k = \frac{\sum_{i,j} Cen_{ijk} P_{DDij} P_{CEi}/P_{Mj}}{\sum_{i,j} Cen_{ijk}}, \tag{7}$$

where P_{CEi} and P_{Mj} are functions of the parameters β_{CEi}, μ_{CEk}, α_{CEik}, θ_{Mjk}, β_{Mj}, μ_{Mk} and α_{Mjk}. (see formulas (1) and (3)). The joint posterior distribution of these parameters, along with Σ and the $\gamma^2_{c_e(i)}$'s and $\gamma^2_{c_p(j)}$'s, is intractable but realizations from the joint posterior were obtained via MCMC methods and inferences were made based on numerical methods.

The prior distribution of each β is specified as an improper distribution consisting of independent uniform, essentially over the entire real line. We follow Gelman's advice (Gelman 2005) and use uniform priors on the square root of all variance parameters. A uniform prior between -1 and 1 was used for the correlation coefficient of Σ.

7. Results

All estimates are based on MCMC moments. Using two different starting values for the hyper parameters, the MCMC appeared to converge to a stationary distribution before the first hundred iterations. As a conservative measure, the first 500 iterations were discarded. A total of 40,000 iterations were

run, each, to obtain estimates for the random effects model and for the synthetic model.

The following two figures summarize the variance component estimates using box plot with endpoints at the 5% and 95% quantiles. Figure 1 provides a summary of the LCO random effects and Figure 2 provides a summary of the model error random effects. As can be seen, although the variance components all indicate a positive effect, the model errors are consistently larger than the LCO effect by more than an order of magnitude. A simple explanation of a single "LCO effect", in addition to the post-strata effects, does not look possible in light of these results. In addition, the covariance term of the LCO effect did not appear to explain any systematic LCO error between the p-and e-sample.

Figure 1: Summary of Posterior Distribution of LCO Error Covariance Components

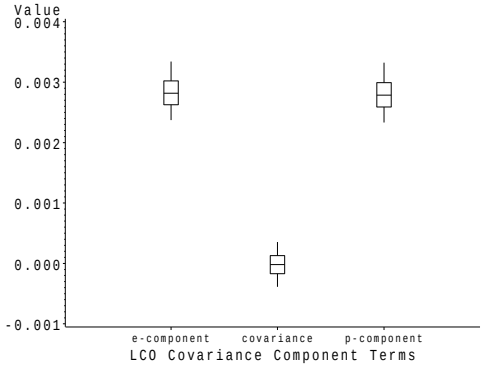
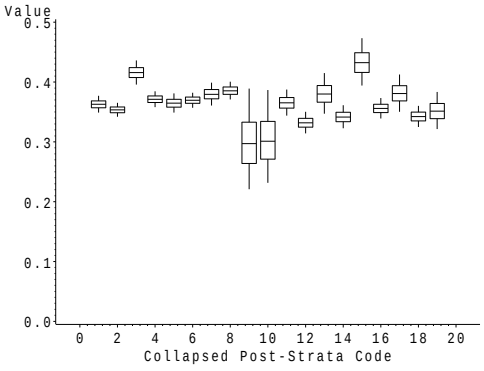


Figure 2: Summary of Posterior Distribution of Model Error Variance Components



The Coverage rate, as seen in (7), is a non-linear function of the Post-strata parameters and the random effects. One way to examine the influence of the components of (7) is to linearize it, in a Taylor series. Expanding around the prior mean of each random effect around its prior mean and each post-strata around their respective average e-sample and average p-sample post strata average one has a linear approximation for the post-strata and random effects.

Table 1: Summary of Post-Strata Variability and Average Model Error Variability

	Post-Strata Variability	Average Model Error Variability
e-sample	0.585	0.361
p-sample	0.395	0.365

Table 1 compares the variability of the linearized post-strata parameters with the variability of the model-errors (LCO effects were not evaluated due to their relatively small size). Both terms appear to contribute equally to the variability of the LCO coverage rate, indicating that the errors not accounted for by post-strata terms may be of equal magnitude as those accounted for. However, disentangling the relationship between the post-strata parameters and the model errors needs to be accomplished before any conclusive statements are made.

Another evaluation of the random effect model on coverage estimates was made by generating values from the posterior distribution of (7) based on the posterior distribution of the parameters from the two models in question. The posterior mean was estimated from the synthetic model, the posterior distribution of the coverage estimates, based on the random effects model, was used to compare how far off the estimates based on the synthetic model was from the random effects model. Note that inference is only on the 521 LCOs with sample. The nineteen other LCOs are small, containing a total of forty-six census records.

Figure 3 exhibits a wide range of differences. However as the 90% probability intervals (based on the shortest, contiguous quantile intervals) indicate in Figure 4, there is quite a lot of error involved. There is still evidence that LCO coverage rates should vary more than indicated by the synthetic model. There were 28 of 521 with intervals that did not cover zero, possibly a chance occurrence. Using a simultaneous rectangular confidence region for all 521 differences, however, we can still state that one LCO interval does not include zero. The confidence intervals are large, indicating that our goal of appending a random effects model to an already large, synthetic model, may not have enough data to draw precise inference about LCO coverage. Another feature of Figure 4 is the fact that the length of the coverage

intervals are only mildly attenuated by sample size, indicating a more complex relationship between the precision of the model components.

Figure 3: Random Effects Model Estimate of Coverage Error: Random Effect Model Estimate - Synthetic Model Estimate

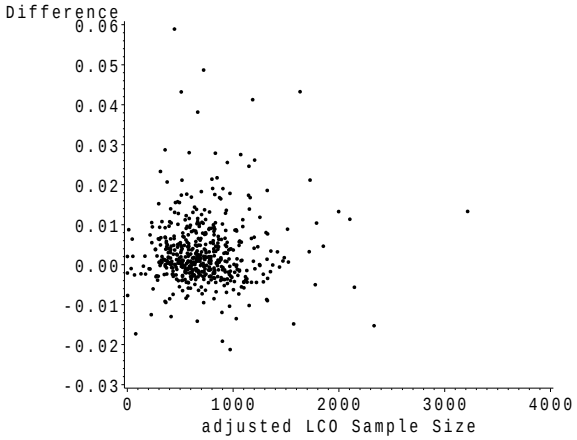
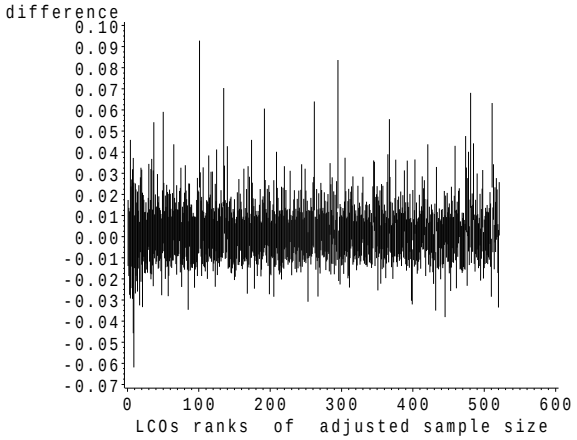


Figure 4: 95% probability intervals of the differences



8. Conclusions

An overall LCO effect did not seem important (based on its estimated variance component) while, surprisingly, individual LCO by grouped post-strata effects exhibited variability comparable to the variability between post-strata. One unanswered question is whether the collapsed-post-strata random effects represent extra small area effect or just represent an attempt of the random effect model to reverse bias caused by the synthetic assumption. Further modeling and reduction of post-strata along with a random effects model may help clear up this question. Another question is whether the collapsed post-strata by LCO effects are, themselves, heterogeneous. Answering this last question through data-analysis may become problematic due to small sample size but, may not be an issue, if a simpler fixed effect model can be shown to replace the current post-strata.

Based on the correctness of the random effects model, it appears that there may be local area effects that are not captured by the synthetic model, as evidenced by estimates of relatively large variance components for LCO by post-strata effects. However, based on the random effects model used, the confidence intervals surrounding small area estimates of LCO coverage are relatively large and do not usually provide for improved prediction of coverage. As intended, this project used random effects to supplement a large fixed-effects model in order to assess the synthetic estimates. If alternative estimates of coverage for small areas are desired, other models could be considered to try to increase overall precision. For example, evaluation of the estimated LCO effects from this project can be used to try to assess other, perhaps more, appropriate levels of geographic variability. The use of fewer fixed effects and further modeling using covariates at various geographic levels (e.g. the LCO level) could be attempted. In addition, more use of statistical modeling principles could be employed. For example, both phase one and phase two design strata could be considered in the model, with a simple multinomial model used to predict the unknown phase two strata identifiers in the population. In addition, modeling correct enumeration as well as e-sample and p-sample captures in a way that more faithfully mirrors the actual process should provide more efficient estimates. Lastly, the within small area scale factors computed using a bootstrap could be based on first conditioning on all covariates in the model, reducing their attenuating effect.

In summary, the use of the random effect model demonstrates that there are important sources of local variation that the synthetic model does not capture (see Figure 2). However, the resulting LCO estimates of coverage were not of good enough precision to strongly argue a change from synthetic (see Figure 4). Future work, will concentrate on using less conservative sample design corrections (to improve precision) and the use of more parsimonious fixed-effects models.

9. Appendix A

The e-sample post-strata were collapsed into 11 groups. This is accomplished by collapsing Tenure, Age, Sex, Date of Return and Household Size categories. In addition, the Race/Hispanic Origin domains[1] used consisted of: in or out of Domain 1: "American Indian or Alaska Native on Reservation", and in or out of Domain 7: "Non-Hispanic or Some other Race". Note that the category "Nuclear" denotes persons in housing units consisting only of the householder along with spouse or own children (17 or younger). "Mailback" (MB) denotes the procedure in which census forms are returned.

Codes for collapsed e-sample post-strata

		Nuclear		Not Nuclear	
proxy	domain	MB	non-MB	MB	MB
yes	1-7	11			
no	1	10		9	
no	2-6	8	7	6	5
no	7	4	3	2	1

Codes for collapsed p-sample post-strata

	Age $\geq$ 18		Age $<$ 18	
domain	not owner	owner	not owner	owner
1-6	19	17	18	16
7	15	13	14	12

10. References

Gelman, Andrew (2005). "Prior distributions for variance parameters in hierarchical models" working paper. http://www.stat.columbia.edu/ gelman/research/published/tau9.pdf

Graubard, Barry I. and Korn, Edward L. (2002) "Inference for superpopulation parameters using sample surveys" Statistical Science, 17, 73-96

Hogan, Howard (1993) "The 1990 Post-Enumeration Survey: Operations and results" Journal of the American Statistical Association, 88, 1047-1060

Hogan, Howard (2003) "The accuracy and coverage evaluation: Theory and design" Survey Methodology, 29, 129-138

Huber, P. J. 1967. "The behavior of maximum likelihood estimates under non-standard conditions" Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 221233.

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

Rubin, Donald B. (1976) "Inference and missing data" Biometrika, 63, 581-590

Seber, G. A. F. (1987) Estimation of animal abundance (Second edition) Charles Griffin & Co (High Wycombe, UK)

Skinner, C. J. (ed.), Holt, D. (ed.) and Smith, T. M. F. (ed.) (1989) Analysis of Complex Surveys. John Wiley & Sons (New York; Chichester

U.S. Census Bureau "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology" (2004). http://www.census.gov/prod/2004pubs/dssd03-dm.pdf

[1] Domain Definitions
1: American Indian or Alaska Native On Reservation
2: American Indian or Alaska Native Off Reservation
3: Hispanic
4: Non-Hispanic Black
5: Native Hawaiian or Pacific Islander
6: Non-Hispanic Asian
7: Non-Hispanic White or "Some Other Race"