Assessing Synthetic Error Via Markov Chain Monte Carlo Techniques

Andrew Keller[1]
[1]U.S. Bureau of the Census, Washington, DC 20233

Abstract
Many small domain estimates require a precise, direct estimate of the within domain variability as one component. However, due to small sample size, the precision of within small domain direct variance estimates of census coverage is questionable. In this paper, Markov chain Monte Carlo (MCMC) techniques are applied to develop a model-based estimate of the within domain variability as part of the estimation process. For this particular application, variability within state is modeled via a random effects model where the census block is the replicate. Ultimately, this block-level model is applied to evaluate synthetic error of the small domain[1].

1. Background

The Accuracy and Coverage Evaluation survey (A.C.E.) used two samples to evaluate coverage for Census 2000, the population sample (P sample) and the enumeration sample (E sample). The P sample used capture/recapture methodology to estimate completeness of the census. The P sample consisted of people rostered from a sample of housing units in a specific location (independent of the census) from a sample of census block clusters (from now on referred to as blocks). It was populated based on the results from a person interview, independent from the census enumerations in the sample blocks.

The E sample estimated the rate of census erroneous enumerations that should not have been included in the census in the sample block cluster or one ring of surrounding blocks. The E sample consisted of census enumerations. It was identified in the same set of census blocks selected for the P sample. E-sample enumerations who matched to P-sample people were counted as correct enumerations. Nonmatched E-sample enumerations underwent a follow up interview to determine whether they were correct enumerations for the specific location.

The A.C.E. divided the population into 416 post-strata where smaller groupings were combined or collapsed to produce more stable estimates. A post-stratum was a group of people sharing demographic and geographic characteristics that were assumed to have similar probabilities of inclusion in the census (U.S. Census Bureau, 2004). A post-stratum, defined at the national level, was the same for both the P and E samples. Within a single post-stratum k, the dual system estimate (DSE) was defined as:

$$DSE_k = census_k \times DDRATE_k \times \frac{CE_k / E_k}{M_k / P_k} \qquad (1)$$

where:

$census_k$: Census count within post-stratum k

$DDRATE_k$: Ratio of data defined census records[2] to all census records within post-stratum k.

[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 author and not necessarily those of the U.S. Census Bureau. The author would like to thank Don Malec for his continued guidance on this project.

[2] In 2000, the census required two characteristics for a record to be data defined. Relationship, sex, race, Hispanic origin, and either age or year of birth counted as characteristics. A valid name also counted as one characteristic. To

CE_k : Weighted estimate of correct census enumerations in post-stratum *k*

E_k : Weighted estimate of census data defined enumerations in post-stratum *k*

M_k : Weighted estimate of matches in post-stratum *k*

P_k : Weighted estimate of P-sample records in post-stratum *k*

The model development described in this paper focuses on the third term for (1), the ratio of the correct enumeration (CE) rate (p_{CE}) and match rate (p_M). The CE rate quantifies the ratio between total E-sample enumerations and a smaller subset of correct E-sample enumerations. The match rate quantifies the ratio between total P-sample people and a smaller subset of P-sample people who matched to a census enumeration. This ratio drives the calculation of (1), and the resulting coverage correction factors (CCF),

$$CCF_k = DDRATE_k \times \frac{p_{CE,k}}{p_{M,k}}.$$

In 2000, census coverage measurement used synthetic techniques to form DSEs for smaller domains. To do this, a census count for the geographic domain of interest g for each post-stratum k was multiplied by the corresponding national-level CCF for that post-stratum. Then, to come up with the DSE for the geographic domain of interest, all the post-strata were summed together:

$$DSE_g = \sum_{\text{all } k} \left[census_{g,k} \times CCF_k \right] \quad (2)$$

If the geographic domain of interest is sufficiently large, then synthetic estimates can constitute suitable coverage estimates. However, as the size of the domain of interest decreases, then the synthetic bias increases.

From the model-based approach used here, CE rate and match rate estimates at the block level can be made. The model-based rates can be used for computation of new CCFs and DSEs at the block level. Although it is possible to model data defined rates as well, that work has been excluded from this model. To come up with estimates for the geographic domain of interest, a calculation similar to (2) is made. However, we now sum across all blocks and post-strata to yield the small domain estimate as shown here:

$$DSE_g = \sum_{\text{all } b} \sum_{\text{all } k} census_{g,b,k} \times CCF_{g,b,k}$$

Traditionally, variance estimates have been computed on large domains where sufficient data was present to calculate a design-based variance of the respective rates. As noted in Malec and Maples (2005), the traditional use of design-based within small domain variance estimates for local coverage has been problematic because there is not a large enough sample. To account for local variation, they develop a model that includes a random effect by local census office (LCO). In their paper, they demonstrate that the synthetic model for coverage does not capture local variation. They suggest development of methods to adequately account for the uncertainty of variability within small domains.

The eventual goal of this model is to assess synthetic error of coverage estimates. To do this, we develop a model for within small domain variance estimation using a random effect at the block level. For this work, the model was applied to selected states by only including each A.C.E. sample block with between 3 and 79 housing units[3].

---

be considered valid by the census, a name had to have at least three characters in the first and last name together. A census record had to be data defined to be eligible for A.C.E. processing.

[3] For A.C.E. sampling, block clusters were classified into four mutually exclusive sampling strata: (a) block clusters with 0 to 2 housing units, (b) block clusters with 3 to 79 housing units, (c) block clusters with 80 or more housing units, and (d) block clusters on American Indian Reservations with three or more housing units. We applied the model only to stratum (b) because strata (a) and (c) are subsampled. Accounting for subsampling in the model would add further complexity that we wanted to avoid during initial research.

2. Methodology

Keller (2007) includes a random effect at the block level to model correct enumeration rate variability within a small domain (a state-level domain in this case). This paper extends that work, adding in modeling of match rate variability and a parameter that correlates the two rates. For consistency in notation, this paper refers primarily to Keller (2007). References to the 2007 work are more frequent in the model development and model checking sections.

2.1 Data Development

In the 2007 paper, we document how the number of correctly enumerated person records and total person records were tabulated for each block in the A.C.E. sample with between 3 and 79 housing units. For this paper, a similar process was followed to create totals for the number of matched person records and total person records at the block-level in the P sample.

2.2 Model Development

Also, in the 2007 work, we derive how the final likelihood (a block-specific binomial distribution) was computed for the E sample, $\prod_{b\in S} L(\mu,\alpha,\varepsilon(b))$

where:

μ : Intercept term

α : Ownership effect

$\varepsilon(b)$: Block effect

In that paper, we further argue the need for an augmented likelihood model based on an assumption that the block effects are normally distributed, $\prod_{b\in S} L(\mu,\alpha,\varepsilon(b))\times\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\varepsilon^2(b)}{2\sigma^2}\right)$.

In this paper, we incorporate the CE and match rates together as a pair of block-specific binomial distributions. We now assume that the block effects have a bivariate normal distribution and a new augmented likelihood model is formed[4]:

$$\prod_{b\in S} L(\mu_*,\alpha_*,\varepsilon_*(b))\times\frac{1}{2\pi\sigma_{CE}\sigma_M\sqrt{1-\rho^2}}\times\exp\left(-\frac{1}{2(1-\rho^2)}\times\left(\frac{\varepsilon_{CE}^2(b)}{\sigma_{CE}^2}+\frac{\varepsilon_M^2(b)}{\sigma_M^2}-\frac{2\rho\varepsilon_{CE}(b)\varepsilon_M(b)}{\sigma_{CE}\sigma_M}\right)\right)$$

where $*$ refers in general to parameters associated with correct enumeration or match rates.

2.3 Model Specifics

Similar to 2007, this analysis used the Metropolis-Hastings algorithm within the Gibbs sampler. It used the Gibbs sampler to draw a single parameter from a conditional distribution given all other parameters. However, since the target distribution was unknown, the Metropolis-Hastings algorithm was used to accept and reject candidates.

For each iteration t, $2B+7$ parameters were processed, where B was the number of random block effects in the model. The remaining seven parameters corresponded to the intercept (μ_{CE},μ_M), ownership effect (α_{CE},α_M), the variance between the block effects $(\sigma_{CE}^2,\sigma_M^2)$, and a parameter indicating the correlation between correct enumeration and match rates (ρ). In our model, we use a non-informative, half-Cauchy prior for σ_* as documented in Gelman (2006). The process cycled through each parameter conditional on the values of the other $2B+6$ parameters and the data by evaluating their Metropolis-Hastings ratios. To properly check for convergence multiple sequences (chains) were run. With respect to following notation, z refers to a chain.

[4] A separate model, with distinct parameters for each state-domain has been made. The state-domain notation, however, has been dropped for ease of reading.

The 2007 paper documents how we sampled our candidate values and processed the Metropolis-Hastings ratios for the block effects, intercept, ownership effect, and variance between the block effects. Since that is unchanged, we omit those explanations from this paper. Below, we document the Metropolis-Hastings ratio for the correlation between correct enumeration and match rates.

2.4 Correlation of Correct Enumeration and Match Rates

A scaled beta distribution was used as the proposal distribution since the correlation was between –1 and 1. For the correlation, we sampled the candidate value by drawing from a beta distribution,

$$\rho(z,t) = 2\xi(z,t) - 1, \xi(z,t) \sim Beta\left[\nu(z,t), \eta(z,t)\right], \nu(z,t) = \frac{B\left(\hat{Q}(z,t)+1\right)}{2}, \eta(z,t) = \frac{B\left(1-\hat{Q}(z,t)\right)}{2}$$

$$\hat{Q}(z,t) = \frac{1}{B-1}\sum_{b=1}^{B}\frac{\left(\varepsilon_{CE}(b,z,t) - \overline{\varepsilon_{CE}(b,z,t)}\right)\left(\varepsilon_{M}(b,z,t) - \overline{\varepsilon_{M}(b,z,t)}\right)}{\hat{\sigma}_{CE}(z,t)\hat{\sigma}_{M}(z,t)}$$

$$\hat{\sigma}_{*}(z,t) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\varepsilon_{*}(b,z,t) - \overline{\varepsilon_{*}(b,z,t)}\right)^{2}}$$

Based on a uniform prior for ρ, its conditional posterior density is:

$$p\left[\rho\left(z,t^{*}\right)\middle|\varepsilon_{*}\left(b=1,z,t\right),...,\varepsilon_{*}\left(b=B,z,t\right),\mu_{*}\left(z,t\right),\alpha_{*}\left(z,t\right),\sigma_{*}\left(z,t\right),y\right]$$

$$\propto \left(1-\rho^{2}\left(z,t^{*}\right)\right)^{-B/2} \times \exp\left(\left[-\frac{1}{2\left(1-\rho^{2}\left(z,t^{*}\right)\right)}\right] \times \left[\frac{\sum_{b=1}^{B}\varepsilon_{CE}^{2}\left(b,z,t\right)}{\sigma_{CE}^{2}\left(z,t\right)} - 2\rho\left(z,t^{*}\right)\left(B-1\right)R(z,t) + \frac{\sum_{b=1}^{B}\varepsilon_{M}^{2}\left(b,z,t\right)}{\sigma_{M}^{2}\left(z,t\right)}\right]\right)$$

$$R(z,t) = \frac{1}{B-1}\frac{\sum_{b=1}^{B}\varepsilon_{CE}(b,z,t)\varepsilon_{M}(b,z,t)}{\sigma_{CE}(z,t)\sigma_{M}(z,t)}$$

The Metropolis-Hastings ratio with respect to the correlation included the proposal function as well as the posterior densities.

2.5 Convergence Analysis

As is done in the 2007 paper, this analysis employed the Gelman and Rubin Method (Gelman, Carlin, Stern, and Rubin, 2000) as its convergence diagnostic. To monitor convergence, we calculated a potential scale reduction factor defined in Gelman et al (2000) for every parameter. For this analysis, the parameter vector subject to convergence monitoring was comprised of the random block effects of each block, the intercept terms, the ownership terms, the variance between the block effect terms, and the term for the correlation between correct enumeration and match rates.

To begin, $z = 10$ starting values for each parameter were chosen as initial values. For those initial values, dispersed starting points were used. This was done to determine if problems existed with the model's convergence and to ensure that the parameter space was thoroughly searched to uncover possible modes. To complete inference, the potential scale reduction factor was calculated at intervals of 100 iterations. As recommended by Gelman et al. (2000), when all parameters had a potential scale reduction factor close to 1, the MCMC method was thought to have converged at that iteration, $t = \tau$.

2.6 Inference and Model Checking

Similar to 2007, after we determined τ, we used the parameter values to calculate modeled correct enumeration and match rates for owners and renters for every iteration between $\tau+1$ and 2τ for each block within each chain. These rates were used as draws from the joint posterior distribution.

In the 2007 paper, we applied posterior predictive checking described in Gelman et al. (2000). We used posterior predictive checking to create new E-sample data from the model. We wanted to check if the new samples were consistent with the observed data for the E sample. For this work, because we model correct enumeration *and* match rates, we implement posterior predictive checking on the ratio between the correct enumeration rate and match rate (from now on, this will be referred to as the CEM ratio) in an identical manner.

To complete posterior predictive checking, we began by creating new E and P samples using binomial trials. Then, we compared coverage intervals for simple means and standard error estimates of the CEM ratios from the new samples to corresponding statistics from the observed sample. The construction of coverage intervals is more thoroughly described in Keller (2007). Tables 1.A and 1.B compare the model-based mean and standard error coverage intervals to the mean and standard error from the observed sample. Recall that, within each state, only a subset of blocks was taken to model CEM ratio variability. Since model results may not be illustrative of the whole state, we refer to them as state-domains in the following tables.

Table 1.A – Correct Enumeration/Match (CEM) Ratio Coverage Intervals

State-Domain	1*	2*	3*	4*	5*	6*	7*	8*
Observed Value	1.0921	1.1483	1.1247	1.0772	1.0703	1.0527	1.0851	1.1131
Coverage Interval Lower Bound	1.0804	1.1147	1.0983	1.0600	1.0555	1.0395	1.0726	1.0979
Coverage Interval Upper Bound	1.1062	1.1831	1.1487	1.0982	1.0851	1.0657	1.0969	1.1341

* - coverage interval covers the observed sample value

Table 1.B – Standard Error of Correct Enumeration/Match (CEM) Ratio Coverage Intervals

State-Domain	1*	2*	3*	4*	5*	6*	7*	8*
Observed Value	0.0123	0.0232	0.0387	0.0137	0.0170	0.0134	0.0129	0.0245
Coverage Interval Lower Bound	0.0106	0.0188	0.0327	0.0118	0.0151	0.0110	0.0116	0.0217
Coverage Interval Upper Bound	0.0141	0.0341	0.0444	0.0189	0.0195	0.0158	0.0149	0.0274

* - coverage interval covers the observed sample value

Table 1.A indicates that the model-based coverage intervals for the CEM ratio cover the observed CEM ratio. Table 1.B shows that the model-based coverage intervals for the standard error of CEM ratio cover the observed standard error of the CEM ratio. Because of this result, we see that our model is able to regenerate the data for the E and P samples and adequately account for block-to-block variation without resorting to design-based methods. Note that the observed values in the table are only from block clusters with 3 to 79 housing units. To follow, we apply the model to form small domain estimates.

3. Results

3.1 Small Domain Estimation

After posterior predictive checking provided promising results that MCMC techniques could be used to model CEM ratios, small domain estimates were generated. To accomplish this, we partitioned the block-level data into two sets. Blocks in sample composed the first set of block-level data. With these blocks, using the parameters from the simulations, we calculated a CE and match rate for every iteration for each sampled block within each chain. Note that a different calculation exists within each block for owners and renters. Because of how we have defined the model, we drop the alpha term when calculating the rates for renters. That is,

$$p_{CE,owners}\left(b,z,t\right)=\frac{e^{\mu_{CE}(z,t)+\alpha_{CE}(z,t)+\varepsilon_{CE}(b,z,t)}}{1+e^{\mu_{CE}(z,t)+\alpha_{CE}(z,t)+\varepsilon_{CE}(b,z,t)}},\ p_{CE,renters}\left(b,z,t\right)=\frac{e^{\mu_{CE}(z,t)+\varepsilon_{CE}(b,z,t)}}{1+e^{\mu_{CE}(z,t)+\varepsilon_{CE}(b,z,t)}}$$

$$p_{M,owners}\left(b,z,t\right)=\frac{e^{\mu_{M}(z,t)+\alpha_{M}(z,t)+\varepsilon_{M}(b,z,t)}}{1+e^{\mu_{M}(z,t)+\alpha_{M}(z,t)+\varepsilon_{M}(b,z,t)}},\ p_{M,renters}\left(b,z,t\right)=\frac{e^{\mu_{M}(z,t)+\varepsilon_{M}(b,z,t)}}{1+e^{\mu_{M}(z,t)+\varepsilon_{M}(b,z,t)}}$$

Blocks not in sample composed the second set of block-level data. For this analysis, blocks not in sample are denoted with b^*. There are generally many more blocks not in sample than blocks in sample. With these blocks, we have no random block effect from the simulations. However, a posterior distribution related to the posterior distribution used in the simulations can be created. To do this, we used a draw from a bivariate normal random variable with mean μ_*, variance σ_*^2, and correlation ρ to generate random block effects for non-sampled blocks. That is,

$$\begin{bmatrix}\varepsilon_{CE}\left(b^*,z,t\right)\\ \varepsilon_{M}\left(b^*,z,t\right)\end{bmatrix}\sim N\left(\begin{bmatrix}\mu_{CE}\left(z,t\right)\\ \mu_{M}\left(z,t\right)\end{bmatrix},\begin{bmatrix}\sigma_{CE}^2\left(z,t\right) & \rho\left(z,t\right)\sigma_{CE}\left(z,t\right)\sigma_{M}\left(z,t\right)\\ \rho\left(z,t\right)\sigma_{CE}\left(z,t\right)\sigma_{M}\left(z,t\right) & \sigma_{M}^2\left(z,t\right)\end{bmatrix}\right)$$

From these random draws, we calculated a correct enumeration and match rate for every iteration for each non-sampled block. That is,

$$p_{CE,owners}\left(b^*,z,t\right)=\frac{e^{\mu_{CE}(z,t)+\alpha_{CE}(z,t)+\varepsilon_{CE}\left(b^*,z,t\right)}}{1+e^{\mu_{CE}(z,t)+\alpha_{CE}(z,t)+\varepsilon_{CE}\left(b^*,z,t\right)}},\ p_{CE,renters}\left(b^*,z,t\right)=\frac{e^{\mu_{CE}(z,t)+\varepsilon_{CE}\left(b^*,z,t\right)}}{1+e^{\mu_{CE}(z,t)+\varepsilon_{CE}\left(b^*,z,t\right)}}$$

$$p_{M,owners}\left(b^*,z,t\right)=\frac{e^{\mu_{M}(z,t)+\alpha_{M}(z,t)+\varepsilon_{M}\left(b^*,z,t\right)}}{1+e^{\mu_{M}(z,t)+\alpha_{M}(z,t)+\varepsilon_{M}\left(b^*,z,t\right)}},\ p_{M,renters}\left(b^*,z,t\right)=\frac{e^{\mu_{M}(z,t)+\varepsilon_{M}\left(b^*,z,t\right)}}{1+e^{\mu_{M}(z,t)+\varepsilon_{M}\left(b^*,z,t\right)}}$$

Using the rates, we calculated dual system estimates for both the sampled and non-sampled blocks:

$$dse_{owners}\left(b,z,t\right)=dd_{owners}\left(b\right)\times\frac{p_{CE,owners}\left(b,z,t\right)}{p_{M,owners}\left(b,z,t\right)},dse_{renters}\left(b,z,t\right)=dd_{renters}\left(b\right)\times\frac{p_{CE,renters}\left(b,z,t\right)}{p_{M,renters}\left(b,z,t\right)}$$

$$dse_{owners}\left(b^*,z,t\right)=dd_{owners}\left(b^*\right)\times\frac{p_{CE,owners}\left(b^*,z,t\right)}{p_{M,owners}\left(b^*,z,t\right)},dse_{renters}\left(b^*,z,t\right)=dd_{renters}\left(b^*\right)\times\frac{p_{CE,renters}\left(b^*,z,t\right)}{p_{M,renters}\left(b^*,z,t\right)}$$

To get an overall estimate across all blocks, we summed up the sampled and non-sampled blocks to get a dual system estimate at each iteration and for each chain:

$$dse\left(z,t\right)=\sum_{b}\left[dse_{owners}\left(b,z,t\right)+dse_{renters}\left(b,z,t\right)\right]+\sum_{b^*}\left[dse_{owners}\left(b^*,z,t\right)+dse_{renters}\left(b^*,z,t\right)\right]$$

3.1.1 Assessing the Model

Schindler (2003) documented results of synthetically-based A.C.E. Revision II (A.C.E. Rev II) estimates for states, counties, and places. However, A.C.E. Rev II estimates were not tabulated by sampling stratum, which was how model-based results were generated. As a result, we had no direct way to compare our model-based results to the synthetic results from 2000. To circumvent this, we developed an ad-hoc means for comparing our small domain model-based estimates to documented results for A.C.E. Rev II. We explain how we created confidence intervals from A.C.E. Rev II and the model below.

Confidence Intervals from A.C.E. Rev II

To derive a confidence interval from the published estimates, we followed a simple scheme. Suppose state A has a census count of 1,000,000 and an A.C.E. Rev II population estimate of 1,010,000 with a standard error of 5,000. To construct a 90 percent undercount confidence interval from A.C.E. Rev II data, we performed these computations:

$$undrct\_itrvl\_ACEII = \left[\frac{(1010000 - (1.645 \times 5000)) - 1000000}{(1010000 - (1.645 \times 5000))}, \frac{(1010000 + (1.645 \times 5000)) - 1000000}{(1010000 + (1.645 \times 5000))} \right] = [0.18\%, 1.79\%]$$

Confidence Intervals from Model

We undertook a similar process to calculate a confidence interval from our small domain model. To save space, iterations between $\tau + 1$ and $\tau + 100$ were used to form the undercount confidence interval from the model instead of iterations between $\tau + 1$ and 2τ. To do this, we first calculated a mean and standard error for all dual system estimates. That is,

$$avg = \frac{1}{1000} \sum_{z=1}^{10} \sum_{t=\tau+1}^{t=\tau+100} dse(z,t), stder = \sqrt{\frac{1}{1000-1} \sum_{z=1}^{10} \sum_{t=\tau+1}^{t=\tau+100} (dse(z,t) - avg)^2}$$

Then, the model-based undercount confidence interval was computed as:

$$undrct\_itrvl\_SAmdl = \left[\frac{avg - (1.645 \times stder) - census}{avg - (1.645 \times stder)}, \frac{avg + (1.645 \times stder) - census}{avg + (1.645 \times stder)} \right]$$

where

$census$: Census count within the sampling stratum for state

Again, it is important to note that confidence intervals from A.C.E. Rev II are based upon data from the entire state. Model-based confidence intervals are based on data only for people who lived in blocks with between 3 and 79 housing units. For 2000, the amount of people who lived in blocks with between 3 and 79 housing units varied by state. In some states, about one-half of the population lived in blocks with between 3 and 79 housing units. In other states, about three-fourths of the population lived in these blocks.

For Table 2, the two confidence intervals are compared. Negative undercount values imply that census overcounted the population according to the model-based estimates. Positive values imply that census undercounted the population according to the model-based estimates.

Table 2 – Comparison of Undercount Confidence Intervals (All totals are percents)

State	Undercount Confidence Interval from A.C.E. Rev II	Undercount Confidence Interval from Model
1	[-0.12 , 0.66]	[0.58 , 2.05]
2	[0.98 , 2.08]	[0.21 , 12.08]
3	[-1.70 , -0.58]	[-1.15 , 8.14]
4	[-1.71 , -0.86]	[-0.05,2.24]
5	[-0.92 , 0.27]	[-0.92,2.50]
6	[-1.85 , -1.04]	[-0.14 , 1.08]
7	[-0.80 , -0.02]	[0.43 , 2.78]
8	[-0.81 , -0.07]	[0.09 , 4.63]

Table 2 indicates two results for the model. First, a wider spread exists within the undercount confidence intervals from the model than within the undercount confidence intervals from A.C.E. Rev II. This result may account for block variability, however the smaller confidence intervals from ACE Rev II are likely due to the use of a larger sample size (the entire state versus only a state-domain for the model-based estimator). It should also be noted that the undercount confidence intervals from A.C.E. Rev II primarily show variability due to sampling. Therefore, the nonsampling errors that can have a major effect on small domain estimates are not reflected in the A.C.E. Rev II confidence intervals. As a result, the A.C.E. Rev II confidence intervals may be larger than indicated.

Second, the estimates of undercounts from the model are generally larger than estimates of undercounts from A.C.E. Rev II. Although this difference may just be due to differences between the state-level and state-domain level undercount rate, it could also be due to the extra-heterogeneity captured in the correlated, CE/match random effects model. In other words, by breaking out the estimates to the block level, it could be that we are eliminating some of the heterogeneity that is present when just post-strata are used and individual blocks are combined. As a result, minimizing this heterogeneity bias causes an increase to the DSE estimates because the CE and match probabilities are positively correlated among blocks. More analysis of the bias properties of the dual system estimator is needed to verify this conjecture.

4. Conclusions and Future Work

This work continues the study of applying MCMC methods to estimate variance of coverage estimates over smaller domains. Ultimately, this approach could be extended to a state-level small domain estimation model, where state-level random effects allow borrowing between states and block-level random effects account for the within state variability.

The limitations of this model should be noted. First, this model has been applied only to a subset of the 2000 A.C.E blocks sampled within each state. It will need to be determined if the inclusion of small blocks with fewer than three households or large blocks with more than 79 households will necessitate a change to the model. Second, the model has only one fixed effect, ownership. Other variables will be incorporated to see how results change. Additionally, a model for data defined rates will be constructed. All of these additions will need to be incorporated to develop a better picture of the model's applicability.

References

Gelman, A. (2006) "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian Analysis*, 1(3):515-533.

Gelman, A.B., Carlin, J.S., Stern, H.S., and Rubin, D.B. (2000) *Bayesian Data Analysis*. Washington D.C.: Chapman and Hall/CRC.

Keller, A. (2007) "Using Markov Chain Monte Carlo for Modeling Correct Enumeration and Match Rate Variability." *Proceedings of the Federal Committee on Statistical Methodology*.

Malec, D. and Maples, J. (2005) "An Evaluation of Synthetic Small Area Census Coverage Error Using a Random Effects Model." *Proceedings of the American Statistical Association, Survey Research Methods Section*, 3355-3362. Alexandria, VA: American Statistical Association.

Schindler, E. (2003) "A.C.E. Revision II - Adjusted Data for States, Counties, and Places." DSSD A.C.E. Revision II Memorandum Series #PP-60.

U.S. Census Bureau. (2004) *Accuracy and Coverage Evaluation of Census 2000: Design and Methodology*.