Using Markov Chain Monte Carlo for Modeling Correct Enumeration and Match Rate Variability

Andrew Keller

U.S. Census Bureau
Washington, DC 20233/andrew.d.keller@census.gov

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.

Abstract

The Census Bureau conducted the Accuracy and Coverage Evaluation (A.C.E.) with the goal of producing estimates of the net coverage error of Census 2000. The A.C.E. used dual system methodology to estimate the net coverage error. Dual system estimates were created for population subgroups called post-strata. A problem in model-based evaluation of coverage with respect to smaller geographies is that the variance between blocks and within those geographies needs to be specified in order to estimate a coverage correction factor. To improve coverage estimates, effort has been made towards advancing models involving smaller geographies. This paper offers Markov chain Monte Carlo (MCMC) methods as a computer intensive method to generate these estimates. Specifically, random effects models of the correct enumeration and match rates at the block level are developed to specify this variance with an accompanying statement of precision.

Background

The 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 estimated persons that should have been enumerated in the census at that location according to census residence rules but were not. 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 census erroneous enumerations that should not have been included anywhere in the census or at the specific location. 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 the same probabilities of inclusion in the census (U.S. Census Bureau 2004). A post-stratum was composed of an E-sample post-stratum and P-sample post-stratum pair. Within a single post-stratum, the dual system estimate (DSE) formula was defined as:

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

where:
k : Post-stratum
$census_k$: The census count within post-stratum k

$DDRATE_k$: The ratio of data defined census records to all census records within post-stratum k[1].

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

E_k : Weighted estimate of 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 Formula 1, the ratio of the correction enumeration (CE) rate and match rate. 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, $CCF_k = \frac{DSE_k}{census_k}$.

The goal of this research is to eventually model CE and match rate variances between blocks within a specified geography. From the model-based approach used here, coverage estimates can be made from the modeled CE rate and match rate. The model-based rates, p_{CE} and p_M, respectively can be used for computation of new DSEs in (2) and coverage corrections factors. That is, for each block b and post-stratum k:

$$DSE_{bk} = census_{bk} \times DDRATE_{bk} \times \frac{p_{CE,bk}}{p_{M,bk}} \qquad (2)$$

The model was applied to selected states by including each A.C.E. sample block with between 3 and 79 housing units. Traditionally, to approximate meaningful variances of the CE and match rates, it has been necessary to move up the geographical hierarchy. That is, variance estimates were computed on larger geographies where sufficient data was present to calculate a design-based variance of the respective rates. By developing the model, the goal was to generate variances with corresponding statements of precision for smaller levels of geography.

Methodology

The model development was geared towards CE rates. In the future, an analogous process will be followed to develop a model for match rates.

Data Development

For the E sample, each census record was assigned a CE probability based on A.C.E. processing (U.S. Census Bureau 2004). To create the data, the CE probability was compared to a uniform number between 0 and 1, $U(0,1)$. Based on that comparison, the record received a binary CE value. For each block in the A.C.E. sample with between 3 and 79 housing units, the number of correctly enumerated person records and total person records were aggregated and a CE rate was generated.

Model Development

This initial model development incorporated between block variability using random effects. Initially, the model used a single fixed effect. Let $p(b,w)$ represent the owner CE rate in a given block b. Similarly, let $p(b,r)$ represent the renter CE rate in a given block b. Since the $p(b,\bullet)$ are sample proportions between 0 and 1, logistic

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

regression was used for the model. Because home ownership has traditionally been a good predictor of correct enumeration, that was included in the logistic model as a fixed effect. Also, since each block has unique characteristics, a random block effect was included in the model. As a result, the owner and renter CE rates were written as $p(b,\bullet) = \dfrac{e^{\mu+\alpha s_i+\varepsilon(b)}}{1+e^{\mu+\alpha s_i+\varepsilon(b)}}$

μ : Intercept term

α : Ownership effect

s_i : Binary variable indicating whether the record i was designated as an owner ($s_i = 1$) or a renter ($s_i = 0$)

$\varepsilon(b)$: Block effect

Recall that each person had a binary CE value. Let $\delta_{i,b,w} = 1$ refer to each correctly enumerated owner record in an arbitrary block b with n_{bw} total owner enumerations. Consequently, $\delta_{i,b,w} = 0$ refers to each erroneously enumerated owner record. Analogously, let $\delta_{i,b,r} = 1$ refer to each correctly enumerated renter record in an arbitrary block b with n_{br} total renter enumerations and $\delta_{i,b,r} = 0$ refer to each erroneously enumerated renter record.

Assuming owner correct enumerations follow a binomial model with a CE rate $p(b,w)$, the likelihood function was written as $L(p(b,w)) = \prod_{i=1}^{n_{bw}} p(b,w)^{\delta_{i,b,w}} (1-p(b,w))^{1-\delta_{i,b,w}}$. Similarly, assuming renter correct enumerations follow a binomial model with a CE rate $p(b,r)$, the likelihood function was written as

$L(p(b,r)) = \prod_{i=n_{bw}+1}^{n_{bw}+n_{br}} p(b,r)^{\delta_{i,b,r}} (1-p(b,r))^{1-\delta_{i,b,r}}$. Combining these likelihoods together,

$$L(p(b,w)) \times L(p(b,r)) = \prod_{i=1}^{n_{bw}} p(b,w)^{\delta_{i,b,w}} (1-p(b,w))^{1-\delta_{i,b,w}} \times \prod_{i=n_{bw}+1}^{n_{bw}+n_{br}} p(b,r)^{\delta_{i,b,r}} (1-p(b,r))^{1-\delta_{i,b,r}} \ . \qquad (3)$$

The CE rates were modeled to involve the parameters of interest: $\mu, \alpha, \varepsilon(b)$. Also, substituting the logistic expression for the respective CE rates into the likelihood function resulted in the following modification to (3):

$$L(\mu,\alpha,\varepsilon(b)) = \prod_{i=1}^{n_{bw}} (\frac{e^{\mu+\alpha s_i+\varepsilon(b)}}{1+e^{\mu+\alpha s_i+\varepsilon(b)}})^{\delta_{i,b,w}} (1-\frac{e^{\mu+\alpha s_i+\varepsilon(b)}}{1+e^{\mu+\alpha s_i+\varepsilon(b)}})^{1-\delta_{i,b,w}} \times \prod_{i=n_{bw}+1}^{n_{bw}+n_{br}} (\frac{e^{\mu+\alpha s_i+\varepsilon(b)}}{1+e^{\mu+\alpha s_i+\varepsilon(b)}})^{\delta_{i,b,r}} (1-\frac{e^{\mu+\alpha s_i+\varepsilon(b)}}{1+e^{\mu+\alpha s_i+\varepsilon(b)}})^{1-\delta_{i,b,r}}$$

By applying what is known about s_i for owners and renters and applying exponential properties, the model was further simplified. Therefore,

$$L(\mu,\alpha,\varepsilon(b)) = \prod_{i=1}^{n_{bw}} (\frac{e^{\mu+\alpha+\varepsilon(b)}}{1+e^{\mu+\alpha+\varepsilon(b)}})^{\delta_{i,b,w}} (\frac{1}{1+e^{\mu+\alpha+\varepsilon(b)}})^{1-\delta_{i,b,w}} \times \prod_{i=n_{bw}+1}^{n_{bw}+n_{br}} (\frac{e^{\mu+\varepsilon(b)}}{1+e^{\mu+\varepsilon(b)}})^{\delta_{i,b,r}} (\frac{1}{1+e^{\mu+\varepsilon(b)}})^{1-\delta_{i,b,r}}$$

$$= (\frac{e^{\mu+\alpha+\varepsilon(b)}}{1+e^{\mu+\alpha+\varepsilon(b)}})^{\sum_{i=1}^{n_{bw}}\delta_{i,b,w}} (\frac{1}{1+e^{\mu+\alpha+\varepsilon(b)}})^{\sum_{i=1}^{n_{bw}}(1-\delta_{i,b,w})} (\frac{e^{\mu+\varepsilon(b)}}{1+e^{\mu+\varepsilon(b)}})^{\sum_{i=n_{bw}+1}^{n_{bw}+n_{br}}\delta_{i,b,r}} (\frac{1}{1+e^{\mu+\varepsilon(b)}})^{\sum_{i=n_{bw}+1}^{n_{bw}+n_{br}}(1-\delta_{i,b,r})}$$

By definition of $\delta_{i,b,\bullet}$,

$$\sum_{i=1}^{n_{bw}} \delta_{i,b,w} = nce_{bw}, \sum_{i=1}^{n_{bw}} (1-\delta_{i,b,w}) = nee_{bw}, \sum_{i=n_{bw}+1}^{n_{bw}+n_{br}} \delta_{i,b,r} = nce_{br}, \sum_{i=n_{bw}+1}^{n_{bw}+n_{br}} (1-\delta_{i,b,r}) = nee_{br}$$

where:

nce_{bw} : Number of correctly enumerated owner records in an arbitrary block b

nee_{bw} : Number of erroneously enumerated owner records in an arbitrary block b

nce_{br} : Number of correctly enumerated renter records in an arbitrary block b

nee_{br} : Number of erroneously enumerated renter records in an arbitrary block b

As a result, the likelihood function for block b was finally written as:

$$L(\mu,\alpha,\varepsilon(b)) = \left(\frac{e^{\mu+\alpha+\varepsilon(b)}}{1+e^{\mu+\alpha+\varepsilon(b)}}\right)^{nce_{bw}} \left(\frac{1}{1+e^{\mu+\alpha+\varepsilon(b)}}\right)^{nee_{bw}} \left(\frac{e^{\mu+\varepsilon(b)}}{1+e^{\mu+\varepsilon(b)}}\right)^{nce_{br}} \left(\frac{1}{1+e^{\mu+\varepsilon(b)}}\right)^{nee_{br}}$$

The final likelihood was: $\prod_{b\in S} L(\mu,\alpha,\varepsilon(b))$. In addition, it was initially assumed that the block effects were normally distributed. This led to an augmented likelihood model, $\prod_{b\in S} L(\mu,\alpha,\varepsilon(b)) \times \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{\varepsilon^2(b)}{2\sigma^2})$. (4)

Markov Chain Monte Carlo

In general, let θ be a vector of unobservable population parameters and y denote observed data. Markov chain Monte Carlo (MCMC) methods iteratively generate dependent samples in the parameter space that converge to a target distribution, $p(\theta \mid y)$.

Bayesian techniques allow inference about θ conditional on the observed data y. Using Bayes' rule, the posterior distribution is $p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y \mid \theta) p(\theta)}{p(y)} \propto p(y \mid \theta) p(\theta)$. The likelihood function from above is another way of writing $p(y \mid \theta)$ since it is proportional to the probability given unknown parameters. The $p(\theta)$ term is called the prior distribution.

Metropolis-Hastings Algorithm

The Metropolis-Hastings Algorithm generates samples from a posterior distribution. The Metropolis-Hastings Algorithm uses an acceptance-rejection scheme to draw samples from a candidate distribution. The algorithm works as follows:

1. Start with an initial value $\theta(t=0)$ where the posterior density is greater than 0.
2. For subsequent iterations $t = 1, 2, ..., T-1, T$; sample a candidate value $\theta(t^*)$ from a candidate distribution, $C(\theta(t^*) \mid \theta(t-1))$. This candidate distribution may not be symmetric.
3. Calculate the Metropolis-Hastings ratio, $Ratio = \frac{p(\theta(t^*) \mid y) \times C(\theta(t-1) \mid \theta(t^*))}{p(\theta(t-1) \mid y) \times C(\theta(t^*) \mid \theta(t-1))}$. If the candidate distribution is symmetric, then the Metropolis-Hastings ratio above is simplified to a ratio of the posterior densities.
4. If $Ratio > U(0,1)$, then $\theta(t^*)$ is accepted and $\theta(t) = \theta(t^*)$. If not, $\theta(t) = \theta(t-1)$. After the candidates are determined to converge to the target distribution, inference is completed.

Gibbs Sampler
For multi-parameter Markov chain applications, the Gibbs sampler is used to cycle through all the parameters. If θ is composed of multiple parameters, then for each iteration, a single parameter θ_g is drawn from a conditional distribution given all other parameters of θ. The Gibbs sampler is a special case of the more general Metropolis-Hastings Algorithm where all candidates are accepted and the target distribution is known.

Model Specifics
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, $B+3$ parameters were processed, where B was the number of random block effects in the model. The remaining three parameters corresponded to the mean (μ), ownership effect (α), and variance between the block effects (σ^2). The process cycled through each parameter conditional on the values of the other $B+2$ parameters and the data by evaluating their Metropolis-Hastings ratios. The σ^2 parameter was the only parameter where a non-constant prior distribution was assumed. Its prior distribution was assumed to be half-Cauchy. To properly check for convergence multiple sequences (chains) were run. With respect to following notation, z refers to a chain.

Block Effects. For the block effect, the candidate value was sampled by drawing from a normal distribution, $\varepsilon(b,z,t^*) \sim N(\varepsilon(b,z,t-1),\sigma^2(z,t-1))$. The normal distribution is symmetric in $\varepsilon(b,z,t^*)$ and $\varepsilon(b,z,t-1)$. As a result, the Metropolis-Hastings ratio with respect to the block effects was simplified to the ratio of posterior densities and the decision to accept was based solely on the ratio of the posterior densities.

Intercept. For the intercept term, the candidate value was sampled by drawing from a normal distribution, $\mu(z,t^*) = \mu(z,t-1) + e$ where $e \sim N(0,1)$. The normal distribution is symmetric in $\mu(z,t^*)$ and $\mu(z,t-1)$. As a result, the Metropolis-Hastings ratio with respect to the intercept was simplified to the ratio of posterior densities and the decision to accept was based solely on the ratio of the posterior densities.

Ownership Effect. For the ownership effect, the candidate value was sampled by drawing from a normal distribution, $\alpha(z,t^*) = \alpha(z,t-1) + e$ where $e \sim N(0,1)$. The normal distribution is symmetric in $\alpha(z,t^*)$ and $\alpha(z,t-1)$. As a result, the Metropolis-Hastings ratio with respect to the ownership effect was simplified to the ratio of posterior densities and the decision to accept was based solely on the ratio of the posterior densities.

Variance Between Block Effects. For the variance between the block effects, the candidate value was sampled by drawing from a gamma distribution,

$$\gamma \sim Gamma[\omega, \psi(z,t)]$$

$$\omega = \lceil (B-1)/2 \rceil$$

$$\psi(z,t) = \frac{1}{2}\sum_{b=1}^{B} \varepsilon^2(b,z,t)$$

$$\sigma^2(z,t^*) = \frac{1}{\gamma}$$

The posterior density for the variance between block effects included a half-Cauchy prior term $\frac{1}{1+\sigma^2(z,t^*)}$. The conditional posterior density was expressed as:

$$p[\sigma(z,t^*) \mid \varepsilon(b=1,z,t),...,\varepsilon(b=B,z,t),\mu(z,t),\alpha(z,t),y] \propto \sigma(z,t^*)^{-\frac{B-1}{2}} \exp[-\frac{\sum_{b=1}^{B}\varepsilon^2(b,z,t)}{2\sigma^2(z,t^*)}]\times\frac{1}{1+\sigma^2(z,t^*)}$$

Since the gamma distribution is not symmetric, then the Metropolis-Hastings ratio with respect to the variance between block effects included the candidate distribution. The candidate distribution was:

$$C(\sigma(z,t^*) \mid \sigma(z,t-1)) \propto \sigma(z,t^*)^{-\frac{B-1}{2}} \exp[-\frac{\sum_{b=1}^{B}\varepsilon^2(b,z,t)}{2\sigma^2(z,t^*)}]$$

The logistic terms had no effect because they had no σ dependence. The $\sigma^{-\frac{B-1}{2}} \exp[-\frac{\sum_{b=1}^{B}\varepsilon^2(b,z,t)}{2\sigma^2(z,t^*)}]$ terms cancelled because they were on opposite sides of the quotient. The simplified Metropolis-Hastings ratio was expressed as:

$$Ratio = \frac{p(\sigma(z,t^*) \mid others)\times C(\sigma(z,t-1) \mid \sigma(z,t^*))}{p(\sigma(z,t-1) \mid others)\times C(\sigma(z,t^*) \mid \sigma(z,t-1))} = \frac{\frac{1}{1+\sigma^2(z,t^*)}}{\frac{1}{1+\sigma^2(z,t-1)}}$$

Convergence Analysis

This analysis employed the Gelman and Rubin Method (Gelman et al. 2000) as its convergence diagnostic. To monitor convergence, a potential scale reduction factor was calculated 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 term, the ownership term, and the variance between the block effects. That is,
$\theta = (\varepsilon(b=1),\varepsilon(b=2),...,\varepsilon(b=B-1),\varepsilon(b=B),\mu,\alpha,\sigma^2)$.

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 one hundred iterations. When all parameters had a potential scale reduction factor close to 1, the MCMC method was thought to have converged at that iteration, $t=\tau$.

Inference

After τ was determined, the parameter values were used to calculate modeled owner and renter correct enumeration totals for every iteration between $\tau+1$ and 2τ for each block within each chain. That is,

$$\begin{aligned} CEmdl_{owners}(b,z,t) &= (\frac{e^{\mu(z,t)+\alpha(z,t)+\varepsilon(b,z,t)}}{1+e^{\mu(z,t)+\alpha(z,t)+\varepsilon(b,z,t)}})\times n_{bw} \\ CEmdl_{renters}(b,z,t) &= (\frac{e^{\mu(z,t)+\varepsilon(b,z,t)}}{1+e^{\mu(z,t)+\varepsilon(b,z,t)}})\times n_{br} \end{aligned} \qquad (5)$$

were used as draws from the joint posterior distribution.

Model Checking

Binomial trials were run to produce new samples. Simple means and standard error estimates from the new samples were compared to corresponding statistics from the observed sample to assess model fit. To do this, for each chain/block/iteration grouping between $\tau+1$ and 2τ, n_{bw} trials were run to get a sample of the number of correctly enumerated owners in that block. Similarly, n_{br} trials were run to get a sample of the number of correctly enumerated renters in that block. The inputted Binomial probabilities were based off the modeled correct enumeration totals for owners and renters in (5). The correct enumeration totals for the binomial trials were computed as follows:

$$CEbin_{owners}(b,z,t) \sim Bin(n_{bw}, p = \frac{CEmdl_{owners}(b,z,t)}{n_{bw}})$$

$$CEbin_{renters}(b,z,t) \sim Bin(n_{br}, p = \frac{CEmdl_{renters}(b,z,t)}{n_{br}}) \quad .$$

$$CEbin(b,z,t) = CEbin_{owners}(b,z,t) + CEbin_{renters}(b,z,t)$$

Next, a mean $(meanCERATEbin)$ and standard error $(seCERATEbin)$ over the blocks were calculated for each iteration between $\tau+1$ and 2τ for each chain. They were calculated as follows:

$$meanCERATEbin(z,t) = \frac{[\sum_{b=1}^{B} CEbin(b,z,t)]}{[\sum_{b=1}^{B} n_{bw} + n_{br}]}$$

$$repCERATEbin(b,z,t) = \frac{[\sum_{b=1}^{B} CEbin(b,z,t)] - CEbin(b,z,t)}{[\sum_{b=1}^{B} n_{bw} + n_{br}] - [n_{bw} + n_{br}]}$$

$$seCERATEbin(z,t) = \sqrt{\frac{B-1}{B}\sum_{b=1}^{B}(repCERATEbin(b,z,t) - meanCERATEbin(z,t))^2}$$

The means and standard errors for all iterations between $\tau+1$ and 2τ for each chain were combined and then sorted from smallest to largest. That resulted in two vectors of size 10τ. That is,

$$\mathbf{meanCERATEvect} = \left[meanCERATEbin(1,\tau+1),...,meanCERATEbin(10,2\tau)\right]$$

$$\mathbf{seCERATEvect} = \left[seCERATEbin(1,\tau+1),...,seCERATEbin(10,2\tau)\right]$$

$$\mathbf{sortedmeanCERATEvect} = sort(\mathbf{meanCERATEvect})$$

$$\mathbf{sortedseCERATEvect} = sort(\mathbf{seCERATEvect})$$

Then, the 5% coverage values were created by taking the mean and standard error for the $0.05 \times 10\tau$ sorted iteration. Similarly, the 95% coverage values were created by taking the mean and standard error for the $0.95 \times 10\tau$ sorted iteration. That resulted in the following coverage intervals:

$$\begin{aligned} &mean\_cvg\_itrvl = [sortedmeanCERATEvect(0.05\times 10\tau), sortedmeanCERATEvect(0.95\times 10\tau)] \\ &se\_\_cvg\_itrvl = [sortedseCERATEvect(0.05\times 10\tau), sortedseCERATEvect(0.95\times 10\tau)] \end{aligned} \qquad (6)$$

As an example, suppose a MCMC model is run for 5000 iterations over 50 blocks with 10 chains. Suppose that, from the Gelman and Rubin Method, the model converges at $\tau = 2000$. For all chains, the means and standard errors for iterations between $\tau+1 = 2001$ and $2\tau = 4000$ are combined and then sorted from smallest to largest. That results in vectors of size 20000 for the mean and standard error. The coverage intervals (6) are then formed by:

$$mean\_cvg\_itrvl = [sortedmeanCERATEvect(0.05 \times 20000), sortedmeanCERATEvect(0.95 \times 20000)]$$

$$se\_\_cvg\_itrvl = [sortedseCERATEvect(0.05 \times 20000), sortedseCERATEvect(0.95 \times 20000)]$$

The coverage intervals were compared to the values observed from the 2000 A.C.E. sample. They were computed as follows:

$$meanCERATEobserved = \frac{[\sum_{b=1}^{B} nce_{bw} + nce_{br}]}{[\sum_{b=1}^{B} n_{bw} + n_{br}]}$$

$$repCERATEobserved(b) = \frac{[\sum_{b=1}^{B} nce_{bw} + nce_{br}] - [nce_{bw} + nce_{br}]}{[\sum_{b=1}^{B} n_{bw} + n_{br}] - [n_{bw} + n_{br}]} \qquad (7)$$

$$seCERATEobserved = \sqrt{\frac{B-1}{B}\sum_{b=1}^{B}(repCERATEobserved(b) - meanCERATEobserved)^2}$$

Results

The goal of this research was to determine the feasibility of applying MCMC techniques to model correct enumeration and (eventually) match rates. To assess this, model-based mean and standard error coverage intervals of the CE rate from (6) were compared to the mean and standard error of the observed sample from (7). One of the key aspects in developing the model was determining if the assumption that the random block effects had a normal distribution from (4) was accurate. The following table sets compare the model-based mean and standard error coverage intervals to the mean and standard error from the observed sample. This section provides results for eight states in each of the four census regions. Recall that, within each state, only a subset of blocks was taken to model CE rate variability. As a result, model results may not be illustrative of the whole state.

Model 1: No Random Block Effect

Initially, the challenge was to justify the inclusion of a random block effect in the model to account for heterogeneity between blocks. To do this, it was necessary to empirically show that omitting a random block effect would not be sufficient to model the observed sample. The results were as follows:

Table 1.A – CE Rate Coverage Intervals

Domain	Observed Value	Coverage Interval Lower Bound	Coverage Interval Upper Bound	Interval Covers Observed Value?
1	0.9476	0.9357	0.9589	Yes
2	0.9413	0.9341	0.9481	Yes
3	0.9722	0.9677	0.9764	Yes
4	0.9447	0.9389	0.9505	Yes
5	0.9243	0.9103	0.9383	Yes
6	0.9449	0.9365	0.9533	Yes
7	0.9369	0.9316	0.9423	Yes
8	0.9519	0.9446	0.9588	Yes

Table 1.B – SE(CE Rate) Coverage Intervals

Domain	Observed Value	Coverage Interval Lower Bound	Coverage Interval Upper Bound	Interval Covers Observed Value?
1	0.0128	0.0037	0.0072	No
2	0.0066	0.0028	0.0043	No
3	0.0033	0.0017	0.0024	No
4	0.0061	0.0023	0.0030	No
5	0.0116	0.0058	0.0108	No
6	0.0072	0.0031	0.0048	No
7	0.0067	0.0025	0.0033	No
8	0.0085	0.0027	0.0039	No

Table 1.A indicates that the model-based coverage intervals for the CE rate cover the observed CE rate. However, Table 1.B shows that the model-based coverage intervals for the standard error consistently underestimate the observed standard error of the CE rate. As a result, it was determined that modeling the random block effect was needed to account for heterogeneity between blocks.

Model 2: Normal Random Block Effect

When it was clear that a random block effect would need to be included, the initial thought was to model the random block effect with a normal distribution. The results were as follows:

Table 2.A – CE Rate Coverage Intervals

Domain	Observed Value	Coverage Interval Lower Bound	Coverage Interval Upper Bound	Interval Covers Observed Value?
1	0.9476	0.9362	0.9579	Yes
2	0.9413	0.9344	0.9479	Yes
3	0.9722	0.9679	0.9763	Yes
4	0.9447	0.9390	0.9500	Yes
5	0.9243	0.9103	0.9378	Yes
6	0.9449	0.9365	0.9530	Yes
7	0.9369	0.9317	0.9421	Yes
8	0.9519	0.9450	0.9584	Yes

Table 2.B – SE(CE Rate) Coverage Intervals

``Domain	Observed Value	Coverage Interval Lower Bound	Coverage Interval Upper Bound	Interval Covers Observed Value?
1	0.0128	0.0096	0.0176	Yes
2	0.0066	0.0058	0.0081	Yes
3	0.0033	0.0026	0.0041	Yes
4	0.0061	0.0054	0.0071	Yes
5	0.0116	0.0084	0.0152	Yes
6	0.0072	0.0060	0.0092	Yes
7	0.0067	0.0059	0.0075	Yes
8	0.0085	0.0067	0.0102	Yes

Table 2.A indicates that the model-based coverage intervals for the CE rate cover the observed CE rate. Table 2.B shows that the model-based coverage intervals for the standard error now cover the observed standard error of the CE rate. Because of the results, it can be inferred that the assumption that the random block effects are normally distributed was correct.

Conclusions and Future Work

This work represents a beginning in studying the feasibility of applying MCMC methods to estimate variance of coverage estimates over smaller geographies. Although the initial results above are promising, this model has only been applied 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 less than three households or large blocks with greater than 79 households will necessitate a change to the model.

As mentioned earlier, future work will include modeling match rates using the same paradigm. Furthermore, since CE and match rates are thought to be correlated within a block, the future model will have to incorporate that relationship. Additionally, a model for data defined rates will be constructed using similar techniques.

References

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

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