Developing an Error Structure in Components of Census Coverage Error

Mary H. Mulry[1] and Bruce D. Spencer
Statistical Research Division, U.S. Census Bureau, Washington, DC
Department of Statistics & Institute for Policy Research, Northwestern University, Evanston, IL

Abstract
The 2010 Census Coverage Measurement Program (CCM) will evaluate the coverage of the 2010 U.S. Census. The 2010 CCM will provide estimates of the components of census coverage error (erroneous enumerations and omissions) separately in addition to estimates of net coverage error. Evaluation studies are underway to examine the quality of the 2010 CCM estimates and provide information for improving census coverage measurement methodology. Synthesizing the results of all the CCM evaluations will aid in forecasting and optimizing tradeoffs among costs and errors for the 2020 census. The current plan is to use a simulation approach in constructing the synthesis and to provide estimates of nonsampling bias in the estimated components of coverage error. This paper explores the use of the evaluation studies to yield estimates of nonsampling error for use in the simulation.

Keywords: census omissions, census erroneous enumerations, net census coverage, nonsampling error

1. Introduction

The U.S. decennial census counts of population are subject to errors known as the *components of census coverage error*, which are omissions and erroneous enumerations. The *net error* is equal to the true population size minus the census count. Estimates of components of coverage error and net errors for the 2010 Census are based on data and analysis from the 2010 Census Coverage Measurement Program (CCM). The number of erroneous enumerations is estimated from validation of a sample of census enumerations, called the *E sample*. The net error is estimated by the difference between the census count and a dual system estimate (DSE) based on the data from both the E sample and the *P sample*, a survey of the household population designed to ascertain inclusion in the census. The E sample and the P sample use the same stratified sample of block clusters. All the census enumerations geographically coded to the sample block clusters, or a subsample of them (in large blocks), are in the E sample. For the P sample, U.S. Census Bureau staff independently constructs a listing of the housing units in the sample block clusters without relying on any of the census addresses. A subsample of the listed addresses may be selected in the large blocks.

This paper describes a plan to synthesize the results of the CCM evaluation studies, assessments, and other studies to develop a better understanding of the error structure in estimates of the components of census coverage error, erroneous enumerations and omissions, and estimates of net census coverage error. There are several goals for the study. One is to assess the combined effect of all the sources of error that can be estimated on the estimates of net census coverage error, erroneous enumerations, and

[1] This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

omissions. Another is to identify the sources that have the largest effect on the estimates of net census coverage error, erroneous enumerations, and omissions. Also, we want to identify the types of nonsampling errors that have the largest effect on the estimates of net census coverage error, erroneous enumerations, and omissions.

2. Background

CCM interviewers list all people living in the sample of housing units for the P sample along with their locations on Census Day and their eligibility for census enumeration. With the data collected, the P sample is matched to the census case-by-case using both computer and clerical operations. When there is uncertainty about whether a P-sample person was actually enumerated in the census (or should have been) or whether an E-sample enumeration should have been included at all or in the sample block, a followup interview collects additional information. Then a final matching operation attempts to make a final resolution. For the estimation of net error, enumerations must have sufficient information (name plus two characteristics) to identify the person with high confidence.

In some cases, nonsampling errors prevent the census enumeration status for P-sample members from being ascertained. For example, insufficient data may prevent identification of the person represented by an enumeration either in the P sample or in the census. There are also practical limitations on how wide a geographic area should be searched for the census enumeration. The dual system estimator is designed so that the limitations balance each other and minimally affect the estimate of net error. However, those limitations complicate both the use of the E sample for estimating the number of enumerations that were erroneous and, in particular, the use of the P sample for estimating the number of omissions. New methodology is being used for the data processing needed to estimate the number of erroneous enumerations for the components of census coverage. The number of omissions is estimated from the difference, true population size minus the number of correct enumerations. Unlike previous census estimates which relied on poststratification for dual system estimation and focused on estimating net error, the current estimates are based on logistic regression models.

The effect of the error structure on the estimates of component errors based on the logistic regression estimator for net coverage error is not well understood. Recent studies have laid the groundwork for this evaluation of the error structure (Mulry 2008, 2009; Spencer 2008, 2009) and are summarized in Mulry and Spencer (2010). These studies of the error structure of the logistic regression estimator for net error have described how various kinds of sampling and nonsampling errors affect the estimates of the net error as well as estimates of omissions and erroneous enumerations. In addition, the studies have provided decompositions of the sampling and nonsampling errors. These decompositions have been useful in designing a schematic plan for using a simulation methodology to synthesize the effect of the sources of error on the estimates of net coverage error, omissions, and erroneous enumerations. The sufficient statistics that will facilitate the simulation of the effect of errors have been identified. Detailed models of errors in the E sample and the P sample have been described.

Studying the error structure of the 2010 CCM estimates of net coverage error, omissions, and erroneous enumerations will enhance the understanding of the accuracy of the 2010 Census and of the accuracy of these estimates of net and component coverage errors in the 2010 Census. The analysis may be used as well for forecasting and optimizing

tradeoffs among costs and errors for the 2020 Census. In addition, a synthesis of information regarding data collection error and data processing error from CPEX studies and other sources will be useful in planning research on coverage measurement methodologies for the 2020 Census. Research on other census coverage measurement methodologies ,as well as on refinements for the CCM methodology, will benefit because they have some overlap in potential error sources.

The error structure for the dual system estimator (DSE) has sources that affect the bias and sources that affect the variance. The sources of error contributing to bias may be classified as errors in the data, ratio-estimator bias, error in the correction for correlation bias, inconsistent recording of characteristics for persons in the P sample and the E sample, error from compensation (imputation and weighting) for missing data, and model error (Alho and Spencer 2005). Other sources of data error contributing to bias include E-sample and P-sample data collection error, P-sample matching error and E-sample processing error, and inconsistent recording of variables used in poststratification or logistic regression modeling (Mulry 2006, Mulry and Spencer 1993). Sources of error that may be viewed as contributing to the variance are sampling error, random data error (response variance), and random error in imputations for missing data.

3. Methodology

The estimate of net error is used in the construction of the estimate of the component error of census omissions, so we must discuss not only the estimation of the components of census coverage error but also the estimation of net error. The plans for estimating the net error use a DSE based on logistic regression estimators for estimating the match rate, the correct-enumeration rate, and the data-defined rate. Those rates are defined from a net error perspective, so that a narrow definition of correct enumeration is used. From those rates, a DSE may be constructed and the net census coverage error can be estimated by subtracting the census count from the DSE. An estimator of erroneous enumerations may be constructed (from the component error perspective) as a weighted sum of the number of erroneous enumerations in the E sample. Then the number of census omissions may be estimated as the difference, the DSE minus the estimated number of correct enumerations.

We turn attention first to the estimation of net error and use of logistic regression modeling. The CCM plans to use the same predictor variables in all three logistic regression models (i.e., for the match rate, the correct-enumeration rate, and the data-defined rate).

To be more specific, we follow Mule (2008, 11) and denote logistic regression estimates of rates by

$\pi_{dd,j}$ predicted data-defined rate for census person j

$\pi_{ce,j}$ predicted correct-enumeration rate for census person j

$\pi_{m,j}$ predicted match rate for census person j.

The predictions are of the form

$$\text{logit}(\pi_{\gamma,j}) = \sum_{i=0}^{I} X_{ij}\beta_{\gamma,i} \qquad (1)$$

with

X_{ij} value of predictor variable i, $i = 0,\ldots,I$, for census or P sample person j,

$\beta_{\gamma,i}$ estimated coefficient for X_{ij} for estimation of rate of status type γ (data-defined, correct-enumeration, or matched).

Mule (2008, 9) notes that "One requirement of the production logistic regression processing is that a standard statistical package like SAS be utilized when running the logistic regressions. These packages allow weights to be utilized when solving the weighted maximum likelihood estimates of the regression coefficients." The use of weighted maximum likelihood (also know as pseudo maximum likelihood) for the logistic regression estimation implies (Alho and Spencer 2005, 119-123) that for each status type γ the estimates $\beta_{\gamma,i}$, $i = 0,\ldots,I$, are a function of $I+1$ sufficient statistics of the form

$$S_{\gamma,i} = \sum_{j} w_j Y_{\gamma,j} X_{ij}$$

with

w_j sampling weight for census person j

$Y_{\gamma,j}$ indicator variable taking the value 1 if census person is of status γ and 0 otherwise.

Thus, the calculations of all 3 sets of $\beta_{\gamma,i}$ depend on the $3(I+1)$ sufficient statistics $S_{\gamma,i}$.

A dual-system estimate of the population in a subgroup (or domain) C is specified to be of the form (Mule 2008, 8-11)

$$N_{0,C} = \sum_{j \in C} PREDSE_j$$

With $PREDSE_j = \pi_{dd,j} \times \pi_{ce,j} / \pi_{m,j}$.

This dual-system estimate does not incorporate an adjustment for correlation bias. The correlation bias adjustment factor for census person j is defined as

$CB_j = c_k$ person j is male and race/age group k

1 otherwise.

with

$$c_k = \frac{\sum_{j \in Female \cap k} PREDSE_j}{\sum_{j \in Male \cap k} PREDSE_j} \times r_{DA,k}$$

and

$r_{DA,k}$ the ratio of males to females in race/age group k as estimated from demographic analysis (DA).

To estimate the total population in domain C one may use a dual-system estimator incorporating a correlation bias adjustment (Mule 2008, 13),

$$DSE_C = \sum_{j \in C} PREDSE_j \times CB_j .$$

Notice that the $3(I+1)$ statistics $S_{\gamma,i}$ not only determine the values of $\beta_{\gamma,i}$ but, along with the census enumeration information and the DA sex ratios $r_{DA,k}$, they determine the values of $N_{0,C}$ and DSE_C.

The preceding discussion pertained primarily to net error. The number of erroneous enumerations (from the component error perspective) may be estimated as a weighted sum of erroneous enumerations. Following the specifications in Mule (2008, 21) but with different notation, the number of erroneous enumerations for domain C may be estimated as

$$EE_C = DD_C \times \frac{\sum_{j \in C} w'_j Y'_j}{\sum_{j \in C} w'_j}$$

With

DD_C data-defined count for domain C

w'_j first-stage ratio-adjusted sampling weight for E-sample case j

Y'_j estimated probability that enumeration j is not a correct enumeration according to the component error definition.

The number of census omissions in domain C may be estimated as

$$Omits_C = DSE_C - CEN_C + EE_C + II_C$$

with II_C equal to whole-person census imputations for domain C and CEN_C equal to the census count for domain C.

To summarize, the estimates of net error, $DSE_C - CEN_C$, and the component estimates of erroneous enumerations, EE_C, and census omissions, $Omits_C$, depend on post-enumeration survey data through the following statistics:

$$S_{\gamma,i}, \quad \sum_{j \in C} w'_j Y'_j, \quad \sum_{j \in C} w'_j, \tag{2}$$

with $i = 0,\ldots,I$ and status type γ referring to data-defined, correct-enumeration, and matched.

These sufficient statistics allow the research strategy to be composed of two pieces, which we will discuss further below:

- The first piece involves modeling the joint error distributions in the sufficient statistics in (2) that are induced by the underlying error structure.

- The second piece is a simulation, involving drawing realizations of the sufficient statistics from their joint distribution and for each realization computing the estimates of interest.

4. Joint error distributions

Part of the study is to develop estimators of the nonsampling errors that can be used in modeling the joint error distributions in the sufficient statistics. The joint distributions will be inputs to the simulation that will produce the distribution of the bias and variance of the net error, omissions, and erroneous enumerations.

Below is a discussion of the approach to estimating the errors, both sampling and nonsampling, and the sources of information available for the different types.

4.1 Errors in data

4.1.1 Sources of error in data

Errors that bias the 2010 CCM estimates may occur during either data collection or data processing. The errors sometimes manifest themselves in different ways in the E sample and P sample but have a commonality. The major CCM data collection phases are independent listing of housing units, housing unit followup (initial and final), person interview, and person followup. Data collection error refers to errors that occur during the creation of the CCM independent list of housing units or during the interaction between the interviewers and the respondents.

The data processing operations have computer and clerical components that are entwined with the data collection operations. The data processing operations have different tasks for the E sample and P sample. Data processing error refers to errors that occur during these tasks. Both the E-sample and the P-sample enumerations undergo computer matching to the entire census and subsequent clerical review of linked pairs to search for E to P matches and E-sample duplicates. The P sample is also matched against itself to search for possible duplicates. For the E sample, when the clerical matching confirms that

the pair is the same person, the matchers also try to determine which the correct enumeration is and which the duplicate is. Errors may affect whether a duplicate enumeration is found and possibly affect the classification of the enumeration in the E sample as correct or erroneous. However, the P sample has both people who report living in the sample blocks on Census Day and people who report moving into the sample blocks between Census Day and the CCM interview day. Errors in identification of census enumerations for the people who say they have not moved may cause the CCM to not follow up and not probe to determine if they really did move. For those who report moving, errors in the identification of enumerations for them during the computerized and clerical searches of all census enumerations may affect whether a matching enumeration is found.

The estimate of erroneous enumerations for components of coverage error is based on different data and definitions than used for estimating the erroneous enumerations for net error, although there is vast overlap. Both estimates of erroneous enumerations – from the component error perspective and from the net error perspective – impact the estimates of the omissions component error.

The processing of the E sample attempts to classify each enumeration into one of the following three categories:

- Correct enumerations that are for people in the housing unit population at their usual residence on Census Day

- Erroneous enumerations that do not represent people in the population at their usual residence on Census Day. The types of erroneous enumerations include

    - Duplicates
    - Enumerations for people not in the U.S. housing unit population (e.g, people who live outside the U.S., and people in group quarters or experiencing homelessness)
    - Enumerations not representing a person in the population (for example, pets, or people born after Census Day or who died before Census Day).

- Enumerations that are not at the person's usual residence on Census Day but are the only enumeration for a person in the housing unit population within the area of interest for estimation, which is nation for this discussion, but could be state, county, or other small area. (Wrong location)

The Census Day residence status of some enumerations may be undetermined. Such cases are coded as unresolved and imputations are made for them in the estimation process.

When an enumeration in the E sample is classified incorrectly, the cause arises from the four basic types of errors listed below with their abbreviations in italics in parentheses:

- errors in identification of duplicate enumerations (*dup*)
- errors in determining membership in the housing unit population on Census Day *(pop*)
- errors in determining the usual residence on Census Day (*ures*)
- errors in the geocoding of the housing unit containing the enumeration (*geo*)

The processing for the P sample attempts to classify persons listed on the Person Interview roster as:

- In the P sample
- Not in the P sample.

Then, for those included the P sample, there is an attempt to find a matching census enumeration so they are classified as

- Match
- Nonmatch

The P-sample inclusion status or the match status of some persons on the roster may be undetermined. Such cases are coded as unresolved and imputations are made for them in the estimation process.

To decide on a person's P-sample status and match status, the P sample determines the following for every person on its roster:

- whether the person is a member of the housing unit population on Census Day
- whether the listing for the person is in the P-sample population
- usual residence on Census Day
- usual residence on CCM Person Interview Day
- whether there is an enumeration in the census at the person's usual residence on Census Day.

When the P-sample inclusion status of a person on the Person Interview roster is classified incorrectly, the cause arises from the four basic types of errors listed below with their abbreviations in italics in parentheses:

- errors in determining membership in the housing unit population on Census Day *(pop*)
- errors in determining the usual residence on Interview Day (ID*ures*)
- errors in determining both the usual residence on Census Day and the usual residence on Interview Day (CDID*ures*)
- errors in the geocoding of the housing unit interviewed (*geo*)

When the match status for a person on the Person Interview roster is classified incorrectly, the cause arises from the three basic types of errors listed below with their abbreviations in italics in parentheses:

- errors in identifying a census enumeration for the person *(cen*)
- errors in determining the usual residence on Census Day (CD*ures*)
- errors in the geocoding of the housing unit interviewed (*geo*)

4.1.2 Evaluations regarding errors in data

Several evaluation studies investigate CCM data collection and processing errors and will provide data for the simulation analysis (Mulry and Adams 2009).

The **Respondent Debriefing (RD)** investigates the errors that occur between the respondent and interviewer regarding the roster of residents, alternate addresses where people could be counted on Census Day, and moves. Experts on residence rules and

CCM procedures will accompany interviewers and debrief respondents after their interview regarding the roster of residents, alternate addresses, and moves.

The **Further Study of CCM Housing Units (FS)** will provide information about errors in geocoding housing units to blocks in the census and in the CCM independent listing. The study will use an extended search to examine the level of error in identifying geocoding errors in the CCM.

The CCM **Recall Bias Study (RBS)** focuses on errors in the reporting of mover status caused by respondent recall error. The delineation between these errors in mover status and those detected in the Respondent Debriefing will need to be made so that errors are not double counted. The CCM Recall Bias Study will link to the Change of Address file which could also provide additional information about data collection error regarding the reporting of moves.

The **Matching Error Study (MES)** will evaluate the level of matching error in the clerical matching operation through an independent rematch of a subsample of the CCM block clusters by an expert matching team. Similar rematch studies evaluated the post-enumeration surveys in 1990 (Davis, Biemer, and Mulry 1992) and 2000 (Bean 2001) and found low levels of matching error. However, the 2010 matching operation has more requirements because of the estimation of components of coverage error and, therefore, is more complicated than previous implementations. This study will not include an extended search, so there will be no overlap with the Further Study of CCM Housing Units for identifying geocoding errors.

The **Administrative Records Study (ARS)** will refine the 2000 duplicate identification methodology that employed administrative records (Mulry et al. 2006) and provided an alternate estimate of census duplicates. In addition, the study will use the administrative records methodology for identifying census duplicates in an alternate identification of census enumerations for persons in the P sample. The ARS may link to the Change of Address file, which could provide additional information about data collection error regarding the reporting of moves. Also under consideration is linking to the Birth Records file which could provide additional information about data collection error regarding the reporting of age. This could be helpful in assessing inconsistency in reports of age in the E sample and P sample. In addition, it could be useful in assessing error in the correlation bias adjustment based on sex ratios from Demographic Analysis since it is age-based.

The **Comparison of Census History with CCM (CCH)** takes a very detailed look at the sequence of census operations and compares results from each operation to CCM. Possible errors in geocoding may be revealed when operations add housing units to the sample block or move housing units from one block to another. A field followup will attempt to confirm or deny the possible geocoding errors.

Table 1 shows the data sources that will be available for estimating the terms of the data collection error and processing error in the E sample caused by different types of errors. Table 2 shows the same information for the P sample. Data collection error is more difficult to estimate for the 2010 CCM than in previous coverage measurement surveys which had evaluation followup studies. The timing of the 2010 CCM would cause an evaluation followup to be conducted after April 1, 2011, which would mean collecting Census Day residency more than a year later, leading to concerns about the accuracy of

reporting. Also, all the probes that were used effectively in past evaluation followups are now in the Person Interview.

A Study of Reasonable Alternatives for Imputation Models that would provide information about error in the model for imputation is not currently planned. Random error from the imputation model fitting may be incorporated into the CCM variance estimates, but estimating error from model selection requires a Study of Reasonable Alternatives for Imputation Models. If such a study is not done, a possible alternative is to use results of the 2000 version of the study to derive estimates. An example would be to assume the ratio of the variance component due to error in the model for imputation to the variance component due to sampling error observed in 2000 also held in 2010. This option is less desirable and would require sensitivity analyses to assess the impact of any assumptions.

There also may be error bounds produced for the 2010 Demographic Analysis estimates that could be useful in assessing error in the correlation bias adjustment based on sex ratios from Demographic Analysis.

Table 1. Sources of Information for E-Sample Error Components

Errors (*causes*)	**Collection Error**	**Processing Error**
Erroneous miscoded Correct; Correct miscoded Erroneous (*dup*, *pop*)	RD	MES ARS
Erroneous miscoded Correct; Correct miscoded Erroneous (*ures*)	RD RBS	MES
Erroneous miscoded Wrong Location; Wrong Location miscoded Erroneous (*pop*)	RD	MES
Erroneous miscoded Wrong Location; Wrong Location miscoded Erroneous (*dup)*	RBS	MES ARS
Correct miscoded Wrong Location; Wrong Location miscoded Correct (*ures*)	RD RBS	MES
Correct miscoded Wrong Location; Wrong Location miscoded Correct (*geo*)	FS CCH	MES

Note: ARS - Administrative Records Study; CCH - Comparison of Census History and CCM Results; FS - Further Study of CCM Missed Housing Units; MES - Matching Error Study; RBS - Recall Bias Study; RD - Respondent Debriefings

Table 2. Sources of Information for P-Sample Error Components

Errors (*causes*)	**Collection Error**	**Processing Error**
Not in P sample miscoded In P sample; In P sample miscoded Not in P sample (*pop, IDures, CDIDures*)	RD, RBS	MES ARS
Not in P sample miscoded In P sample; In P sample miscoded Not in P sample (*geo*)	FS CCH	MES ARS
Match miscoded Nonmatch; Nonmatch miscoded Match (*CDures*)	RD RBS	MES ARS
Match miscoded Nonmatch Nonmatch miscoded Match (*cen*)	RD	MES ARS
Match miscoded Nonmatch; Nonmatch miscoded Match (*geo*)	FS CCH	MES

Note: ARS - Administrative Records Study; CCH - Comparison of Census History and CCM Results; FS - Further Study of CCM Missed Housing Units; MES - Matching Error Study; RBS - Recall Bias Study; RD - Respondent Debriefings

4.2 Error from Inconsistent Classification

Inconsistent reporting of variables in the E sample and P sample may cause a bias if they are covariates in the logistic regression models for the DSE. Such a bias will then affect the estimates of net coverage error and possibly omissions. Such bias occurs when the prediction model is fitted to data where the covariates are measured in the P sample, but the prediction model is applied to covariates as measured in the census, which is how the estimates of net error will be constructed (Mule 2008). To measure the impact of this bias on net coverage error estimate, the strategy is to calculate a P-sample match rate corrected for the error due to inconsistently reported covariates, and then use this corrected P-sample match rate to compute a DSE estimate adjusted for inconsistency bias. A comparison of the DSE with a DSE adjusted for inconsistency will produce an estimated bias term.

4.3 Error from Missing Data

CCM data may be missing for a variety of reasons – some CCM interviews fail to take place, some households provide incomplete data on questionnaire items, and sometimes the information for classification as a match or nonmatch is ambiguous. Incomplete and ambiguous data in the E sample can also result in not being able to classify census enumerations as correct or erroneous. The CCM estimation program selects methods for compensating for the different types of nonresponse.

The planned simulation approach for synthesizing errors will model missing data as a variance component. Estimating error from model selection requires a Study of Reasonable Alternatives for Imputation Models (Kearney 2002).

4.4 Sampling Error

Sampling error gives rise to random error, quantified by sampling variance, and to a systematic error known as ratio-estimator bias, which arises because, even if X and Y are unbiased, X/Y typically is biased. The DSE is just such a ratio so the DSE could be biased. Therefore, the estimates of net error and omissions also could be biased. The replication methodology used to estimate the random sampling error also can be programmed to provide an estimate of the ratio-estimator bias.

4.5 Error in the Correlation Bias Correction

Correlation bias is the error in dual system estimation that arises because of a violation of the assumption of independence between the census and the P Sample or because of a violation of the assumption that the enumeration probabilities are equal. Correlation bias tends to be a source of downward bias in dual system estimates. The Census Bureau attempts to preserve the independence of the census and P Sample by keeping the CCM data collection and processing operations completely separate from the census data collection and processing. The logistic regression modeling groups the respondents by geography, sex, age, racial and ethnic groups, and population density and thereby reduces the bias by grouping together people with similar chances of being counted, as estimated by the match rate. This approach was first recommended by Chandrasekar and Deming (1949) using poststratification. However, the groupings used by logistic regression modeling or poststratification may not describe all the heterogeneity of enumeration probabilities and thereby may not eliminate all correlation bias; see Section 3.

Corrections for correlation bias in dual system estimates for adult males have been developed using Demographic Analysis estimates of the sex ratios (the ratios of the number of males to the number of females) (Bell 1993). Demographic Analysis is the only method viewed as producing estimates of quality high enough for this correction. The CCM plans to employ the two-group model for the correlation bias correction, which is the same model used in the construction of the Accuracy and Coverage Evaluation Revision II estimate (U.S. Census Bureau 2003). The application of the two-group model uses two racial groups, black and nonblack, by age because the historical records that Demographic Analysis uses contain only these groups.

If there is information about the error in the correlation bias correction from studies associated with CCM or Demographic Analysis, this information will be incorporated into the simulation to synthesize the errors. If no information about error in the correlation bias correction is available, then sensitivity analyses may be done.

4.6 Estimators of Nonsampling Errors

Although models of errors in the E sample and the P sample exist (Mulry and Spencer 2010), there are open research questions regarding the design of E-sample and P-sample nonsampling error estimation when using such a wide variety of error components. The design of estimators for the E-sample and P-sample errors that are suitable for adjusting the sufficient statistics in the simulation requires care and is not simple. There is no standard way to determine how an error source affects the sufficient statistics. Each sufficient statistic is affected in a different way by an error source.

A further complication is caused by the type of data that will be available concerning the sources of error. Several data sources with independent collection methods will have to be examined in the course of designing the estimators. The CCM evaluation studies producing the data are shown in Tables 1 and 2, as well as the analyses conducted during the course of the construction of the CCM estimates. The design of the nonsampling error estimation must estimate each error separately in a manner that does not double count errors, but does account for correlation between errors.

Estimates of bias and variance components for data errors need to be developed from the CCM evaluations and the CCM Recall Bias Study as indicated in Tables 1 and 2. The vector of nonsampling bias components needs to be estimated as well as the covariance matrix for the nonsampling error components. The structure of these moment specifications will somewhat parallel those for the evaluation of the Census 2000 Accuracy and Coverage Evaluation Study, as described in Mulry and ZuWallack (2002).

5. Simulation

The simulation design relies on the fact that each of the sufficient statistics is a weighted sum or total, which simplifies estimation of the effects of nonsampling error. This is the case for the estimates of net coverage error, omissions, and erroneous enumerations. Once the nonsampling errors, their variances, and covariance matrix are estimated, the simulation will draw repeatedly and independently from their joint distribution to produce the distribution of a bias estimate. For computing purposes, we intend to use a logistic regression package that offers the option to input sufficient statistics directly rather than compute the sufficient statistics from the individual data records. Alternatively, but more tediously, the data could be reweighted (by raking) to match the revised sets of sufficient statistics.

We will apply the probability models for errors to simulate the joint distribution for the sufficient statistics in (2). The probability distributions will be centered on the observed values adjusted for the estimated biases, and their random component will be derived from a multivariate normal specification with mean vector equal to zero and covariance matrix. Simulation from this distribution will yield distributions for the statistics in (2), from which we can derive distributions for the estimates of net error, $DSE_C - CEN_C$, and the component estimates of erroneous enumerations, EE_C, and census omissions, $Omits_C$, for domains C to be chosen as part of the research. Differences between the means of the latter distributions and the original estimates indicate the estimated biases in the original estimates, and the standard deviations indicate the standard deviations of the sampling distributions.

As shown in the discussion of the joint error distribution, the probability models will be developed somewhat differently for (1) sampling error, (2) error from missing data, (3) effect of inconsistent classification, (4) other data errors and processing error, and (5) error in the correlation bias adjustment.

6. Analysis of results

Part of the research will be to specify the domains C for the analysis. For each domain, and for the estimates of net error, $DSE_C - CEN_C$, and the component estimates of erroneous enumerations, EE_C, and census omissions, $Omits_C$, the following statistics will be computed from the simulated distribution: (i) estimate of bias, (ii) estimate of standard deviation (reflecting both sampling error and random nonsampling errors), and (iii) deciles of the distribution.

The analyses will include a sensitivity analysis to aid in determining the most influential error sources and error types.

References

Alho, J. M. and Spencer, B. D. (2005) *Statistical Demography and Forecasting.* Springer. New York, NY.

Bean, S. L. (2001) "ESCAP II: Accuracy and Coverage Evaluation Matching Error." Executive Steering Committee For A.C.E. Policy II, Report No. 7. October 12, 2001. U.S. Census Bureau, Washington, DC.

Bench, K. (2002) "Evaluation Report for P-sample Match Rate Corrected for Error due to Inconsistent Post-stratification Variables." DSSD A.C.E. Revision II Memorandum Series # PP-46. PRED Census And Survey Measurement Staff Memorandum Series: CSM-ACE-REVISION II-R3R. December 31, 2002. Washington, D.C.: U.S. Census Bureau.

Chandrasekar, C. and Deming, W. E. (1949) "On a Method of Estimating Birth and Death Rates and the Extent of Registration". *Journal of the American Statistical Association*, 44, 101-115.

Bell, W. R. (1993) "Using Information from Demographic Analysis in Post-Enumeration Survey Estimation." *Journal of the American Statistical Association*, 88, 1106-1118.

Davis, M., Biemer, P. and Mulry. M. H. (1992) "Matching Error Study." *Proceedings of the Survey Research Methods Section.* American Statistical Association. Alexandria, VA. 170-175.

Kearney, A. T. (2002) A.C.E. Revision II Missing Data Evaluation. DSSD A.C.E. Revision II Memorandum Series #PP-48. PRED Census And Survey Measurement Staff Memorandum Series: CSMACE-REVISION II-R2R. December 31, 2002. Washington, D.C.: U.S. Census Bureau.

Mule, T. (2008) 2010 Census Coverage Measurement Estimation Methodology. DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-E-18. October 30, 2008. Washington, D.C.: U.S. Census Bureau.

Mulry, M. H. (2009) A Study of Sources for the Error Structure in Estimates of Census Coverage Error Components. *Proceedings of the 2009 International Total Survey Error Workshop*. National Institute for Statistical Sciences. Research Triangle Park, NC.

Mulry, M. H. (2008) "Error Structure in Estimates of Census Coverage Error Components." *Proceedings of the 2008 International Total Survey Error Workshop*. National Institute for Statistical Sciences. Research Triangle Park, NC.

Mulry, M. H. (2006) “Summary of Accuracy and Coverage Evaluation for Census 2000." *Journal of Official Statistics*. 23, 345-370.

Mulry, M. H. and Adams, T. S. (2009) “Overview of Evaluations of the 2010 Census Coverage Measurement Program.” *JSM Proceedings, Survey Research Methods Section.* American Statistical Association. Alexandria, VA. 3117 – 3128.

Mulry, M. H., Bean, S. L., Bauder, D. M., Wagner, D., Mule, T. and Petroni, R. (2006) “Evaluation of Estimates of Census Duplication Using Administrative Records Information.” *Journal of Official Statistics*. 22. Statistics Sweden. Stockholm, Sweden. 655 – 679.

Mulry, M. H. and Spencer, B. D. (2010) “The Structure of Sampling and Nonsampling Error Components in 2010 Census Coverage Error Estimation.” Unpublished manuscript. Statistical Research Division. U.S. Census Bureau. Washington, DC.

Mulry, M. H. and Spencer, B. D. (1993) “Accuracy of the 1990 Census and Undercount Adjustments.” *Journal of the American Statistical Association*, 88, 1080-1091.

Mulry, M. H. and ZuWallack, R. (2002) A.C.E. Revision II Confidence Intervals and Loss Functions. DSSD A.C.E. Revision II Memorandum Series #PP-42. December 31, 2002. Washington, D.C.: U.S. Census Bureau.

Spencer, B. D. (2009) Research Designs for Investigating the Error Structure in Net Census Coverage Error Modeling. Report for the U.S. Census Bureau prepared under contract YA1323-09-SE-0174. U.S. Census Bureau. Washington, D.C.

Spencer, B. D. (2008) Investigation of Errors in Direct Estimates Used for Validation of Variable Selection for Net Census Coverage Error Modeling. Report for the U.S. Census Bureau prepared under contract YA132306SE0513. U.S. Census Bureau. Washington, D.C.

U.S. Census Bureau (2003) “Technical Assessment of A.C.E. Revision II.” March 12,2003. U.S. Census Bureau. Washington, DC. http://www.census.gov/dmd/www/pdf/ACETechAssess.pdf