Designing Estimators of Nonsampling Errors in Estimates of Components of Census Coverage Error

Mary H. Mulry[1] and Bruce D. Spencer
Center for Statistical Research & Methodology, 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) plans to use logistic regression models for the correct enumeration rate, data-defined rate, and match rate that will be used in dual system estimation of the population size which then will be used to measure net census coverage error. Recent 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 study the combined effect of the sources of error on the estimates of net coverage error as well as on the estimates of omissions and erroneous enumerations. The sufficient statistics that will facilitate the simulation of the effect of errors have been identified. This paper discusses the design of estimators of nonsampling errors suitable for use with data sources that will be available from the CCM and the evaluations of the CCM.

Key words: census omissions, census erroneous enumerations, undercount

1. Introduction

A key objective for the 2010 Census Coverage Measurement (CCM) program is to obtain separate estimates of erroneous census inclusions and census omissions. Producing estimates of net error for the 2010 census continues to be an important objective. The plan for estimating census omissions is to sum estimates of net coverage error and erroneous enumerations. In addition the plans include estimating the net coverage error for demographic groups and by geography (Cantwell and Ramos 2010).

Likewise, the evaluations of the CCM share the goal of providing information useful for designing improvements in census-taking methodology and coverage measurement methodology. In addition, as part of the 2010 Census Program for Evaluations and Experiments (CPEX), the CCM evaluations have the goal of providing information about the quality of the CCM operations and estimates.

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. The analysis may be used as well in the planning of the 2020 Census. In addition, a synthesis of information regarding data collection error and data processing error studies 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.

[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.

For 2010, the CCM plans to use logistic regression modeling in the estimation of net census coverage error rather than poststratification, the approach used for previous censuses. Logistic regression models for the correct enumeration rate, data-defined rate, and match rate will be used in dual system estimation (DSE) of the population size which then would be used to measure net census coverage error.

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 of the error structure of the logistic regression estimator for net error have described how, via formulas, various kinds of sampling and nonsampling errors affect the estimates of the net error, omissions, and erroneous enumerations Mulry and Spencer (2010a, 2010b). In addition, the studies have provided decompositions of the sampling and non-sampling 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 as well as on the estimates of 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.

However, the design of estimators for the E- and P-sample errors that are suitable for use with the sufficient statistics in the simulation is complicated. A further complication is caused by the type of data that will be available concerning the sources of error. Several data sources will have to be examined in the course of designing the estimators. The CPEX studies producing the data include the Matching Error Study (MES), the Respondent Debriefings (RD), the Further Study of CCM Missed Housing Units (FS), the Comparison of Census History Study (CCH) also known as the Processing History Study, the Evaluation of Address Frame Accuracy and Quality, and the CCM Recall Bias Study (RBS). Other sources of information about E- and P-sample errors are the assessments conducted during the course of the construction of the CCM estimates.

In this paper we discuss the general form of the estimators of nonsampling errors and the data available to aid in forming the estimates of nonsampling errors. Section 2 has a background discussion on the CCM data collection and estimation. Sections 3, 4, and 5 contain the form of the estimators of nonsampling error. Section 6 reviews the evaluation studies that will be carried out to understand their design both in terms of selection of cases and variables measured. Also there is a discussion in Section 7 of how the design of the evaluation studies relates to the estimation of moments of distributions of the nonsampling errors. Finally we close in Section 8 with a summary and discussion of future work.

2. CCM Background

2.1 CCM Data

The CCM sample is composed of two overlapping samples, the enumeration sample (E sample) and the population sample (P sample) which is conducted independently of the census. CCM interviewers attempt to conduct a personal interview (PI) at each housing unit selected in the P sample to list all people living there 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, the enumeration 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. E.g., 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 clerical data processing needed to estimate the number of erroneous enumerations for the components of census coverage.

2.2. CCM Estimation

For the DSE, CCM fits three logistic regression models,

$$\log[\hat{\pi}_{\gamma,j}/(1-\hat{\pi}_{\gamma,j})]=\sum_{i=0}^{I}X_{ij}\hat{\beta}_{\gamma,i}$$

where

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

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

$\gamma = DD, CE, M$ corresponding to data-defined, correct-enumeration, and match.

The data-defined are assumed to be identified without error, so we are concerned about errors in identifying the statuses *CE* and *M*. The sufficient statistics are of the form

$$S_{\gamma,i}=\sum_{j}w_jY_{\gamma,j}X_{ij}.$$

with

w_j sampling weight for census person j

$Y_{\gamma,j}$ indicator variable equal to 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

$$CB_j=\begin{cases}0 & \text{when person } j \text{ is male and race/age-group } k\\ 1 & \text{otherwise}\end{cases}$$

with $c_k=\dfrac{\sum\limits_{j\in Female\cap k}PREDSE_j}{\sum\limits_{j\in Male\cap k}PREDSE_j}\times r_{DA,k}$

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),

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

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

$$EE_{comp-C} = DD_C \frac{\sum_{j \in C} w_j^{'} Y_j^{'}}{\sum_{j \in C} w_j^{'}}$$

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

$$Omits_C = DSE_C - CEN_C + EE_{Comp-C} + II_C$$

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

We want to estimate the effect of errors in the statuses CE and M on the DSE. We will do this by adjusting sufficient statistics to account for systematic nonsampling errors; specifically, we estimate biases in the statistics and then subtract the bias estimates from the statistics. The bias estimates will come from estimated error rates for cases in various operationally defined categories.

3. Error in Matches

Persons listed on the rosters collected in the CCM Person Interview (PI) receive both a residency status code and a match code. The residency code indicates whether the person should be in the P sample and if so, whether the person is (1) a nonmover who lived in the sample block cluster on Census Day and PI interview day, (2) a mover who moved from a housing unit into the sample block cluster between Census Day and PI interview day, or (3) an outmover who lived in sample block cluster on Census Day and moved to a group quarters or out of the country before the PI interview.

To simplify for this paper, we are assuming that there are no errors in the classification of whether people are validly in the P-sample. Their residency status as a nonmover, mover, or outmover may be in error, and if so, it potentially only affects their match status because CCM may search in the wrong location for their census enumeration.

3.1. Definition of Error Rates for Matches

Suppose we know have estimates of the two types of error rates for those who match a census enumeration at their Census Day residence and receive a match status M. The two error rates are the rate nonmatches are miscoded as matches and the rate that matches are miscoded as nonmatches. We have these estimates for the following six categories:

- Nonmovers who did not go to followup
    - r_{1mn} = rate nonmatches miscoded as matches
    - r_{1nm} = rate matches miscoded as nonmatches
- Nonmovers who went to followup
    - r_{2mn} = rate nonmatches miscoded as matches
    - r_{2nm} = rate matches miscoded as nonmatches
- Movers who did not go to followup
    - r_{3mn} = rate nonmatches miscoded as matches
    - r_{3nm} = rate matches miscoded as nonmatches
- Movers who did not go to followup
    - r_{4mn} = rate nonmatches miscoded as matches
    - r_{4nm} = rate matches miscoded as nonmatches
- Outmovers who did not go to followup
    - r_{5mn} = rate nonmatches miscoded as matches
    - r_{5nm} = rate matches miscoded as nonmatches
- Outmovers who went to followup.
    - r_{6mn} = rate nonmatches miscoded as matches
    - r_{6nm} = rate matches miscoded as nonmatches.

If more detailed data were available from the evaluation studies, then instead of using simple misclassification rates, it could be possible to use logistic regression modeling to account for variability in the rates based on the available covariates in the evaluation studies and the CCM sample (here, the P sample). The latter approach is more complex but is receiving attention in the literature (e.g., Lyles et al. 2011; Katz and Katz 2010; Carroll, Ruppert, Stefanski, and Crainiceanu 2006; Paulino, Soares, and Neuhaus 2003; Neuhaus 2002; Magder and Hughes 1997), where joint analysis of the main sample and the validation sample is considered. A related approach already used by the Census Bureau for adjustment for nonsampling error in the DSE for Census 2000 is to use double sampling (U.S. Census Bureau 2003).

3.2. Error Components for Matches

Let us take one explanatory variable, say X_{1j} = 1 if *j* is male and 0 if *j* is female. We have the following weighted estimates of males (*i*=1) by matches and nonmatches in the categories above:

- Nonmovers who did not go to followup
    - M_{11} = matches
    - N_{11} = nonmatches
- Nonmovers who went to followup
    - M_{12} = matches
    - N_{12} = nonmatches
- Movers who did not go to followup
    - M_{13} = matches
    - N_{13} = nonmatches
- Movers who did not go to followup
    - M_{14} = matches
    - N_{14} = nonmatches
- Outmovers who did not go to followup
    - M_{15} = matches

  - N_{15} = nonmatches
- Outmovers who went to followup.
  - M_{16} = matches
  - N_{16} = nonmatches.

The estimated errors in the sufficient statistic for i=1 (Male) by the categories would be as follows:

- Nonmovers who did not go to followup:
  $e_{11} = (r_{1mn}\, N_{11} - r_{1nm}\, M_{11})/(1 - r_{1mn} - r_{1nm})$
- Nonmovers who went to followup
  $e_{12} = (r_{2mn}\, N_{12} - r_{2nm}\, M_{12})/(1 - r_{2mn} - r_{2nm})$
- Movers who did not go to followup
  $e_{13} = (r_{3mn}\, N_{13} - r_{3nm}\, M_{13})/(1 - r_{3mn} - r_{3nm})$
- Movers who did not go to followup
  $e_{14} = (r_{4mn}\, N_{14} - r_{4nm}\, M_{14})/(1 - r_{4mn} - r_{4nm})$
- Outmovers who did not go to followup
  $e_{15} = (r_{5mn}\, N_{15} - r_{5nm}\, M_{15})/(1 - r_{5mn} - r_{5nm})$
- Outmovers who went to followup
  $e_{16} = (r_{6mn}\, N_{16} - r_{6nm}\, M_{16})/(1 - r_{6mn} - r_{6nm})$.

The divisor, $1 - r_{jmn} - r_{jnm}$, is needed because the true numbers are unknown, and are being derived as sums of (differently) weighted fractions of *observed* numbers.

The derivation of the estimators of the net errors uses the so-called "inverse matrix method" (Morrissey and Spiegelman 1999). Then the corrected sufficient statistic for status $\gamma = M$ corresponding to $i = 1$ has the form

$$S_{M,1} = \sum_j w_j Y_{M,j} X_{ij} - \left(e_{11} + e_{12} + e_{13} + e_{14} + e_{15} + e_{16}\right).$$

Estimation of error rates is discussed in Section 7.1.

4. E-Sample Errors from the Net Error Perspective

Now we turn our attention to errors in whether census enumerations in the E sample are coded as correct (CE) or erroneous (EE). Estimating net error requires processing of the E sample from a different perspective than required for component error estimation. Census enumerations that do not have at least two characteristics are considered to not be data defined and all the characteristics are imputed. The E sample contains only data-defined enumerations. To be used in the estimation of net error, an enumeration in the E sample has to have sufficient information for the matching operation to identify the person, viz. a name and two characteristics. The estimation of erroneous enumerations from the component error perspective includes all the enumerations in the E sample.

4.1. Definition of Error Rates

To define error rates for the enumeration status CE, we have an analysis for males (i=1) similar to Section 3.1. Continue to consider X_{1j} = 1 if j is male and 0 if j is female. We have the following weighted numbers of enumerations of males by enumeration status (correct and in the correct location, erroneous, in the wrong location but otherwise correct). It is also convenient to introduce notation for data-defined cases with insufficient information for matching. The weighted numbers are based on the E sample and are restricted to data defined cases only. Whether a case has sufficient information for matching and whether it went to followup are known without error. We have the following weighted estimates of males (i=1) by correct and not correct enumeration status from the net error perspective:

- Enumerations with sufficient information for matching that did not go to followup
  - CE_{11} = correct enumeration in the correct location
  - NC_{11} = not correct enumeration in the correct location
    - $NC_{11} = EE_{11} + WL_{11}$
    - EE_{11} = erroneous enumeration
    - WL_{11} = in the wrong location but otherwise correct enumeration
- Enumerations with sufficient information for matching that did go to followup
  - CE_{12} = correct enumeration in the correct location
  - NC_{12} = not correct enumeration in the correct location
    - $NC_{12} = EE_{12} + WL_{12}$
    - EE_{12} = erroneous enumeration
    - WL_{12} = in the wrong location but otherwise correct enumeration.

4.2. Error Components

Suppose we have estimates of the two types of error rates for status CE, which are the rate correct enumerations in the correct location (CE) are miscoded as other than that (NCE) and the rate not-correct enumerations (NCE) are miscoded as correct enumerations in the correct location (CE). We have these estimates for the following two categories of enumerations.

- Enumerations with sufficient information for matching that did not go to followup
  - r_{1cn} = rate NC miscoded as CE
  - r_{1nc} = rate CE miscoded as NC
- Enumerations with sufficient information for matching that did go to followup
  - r_{2cn} = rate NC miscoded as CE
  - r_{2nc} = rate CE miscoded as NC

The estimated errors in the sufficient statistic for $i=1$ (Male) by the categories follow:

- Cases with sufficient information for matching that did not go to followup

$$e_{11} = \left(r_{1cn} NC_{11} - r_{1nc} CE_{11}\right) / \left(1 - r_{1cn} - r_{1nc}\right)$$

- Cases with sufficient information for matching that did go to followup

$$e_{12} = \left(r_{2cn} NC_{12} - r_{2nc} CE_{12}\right) / \left(1 - r_{2cn} - r_{2nc}\right)$$

Then the corrected sufficient statistic for status γ = *CE_net* corresponding to i = 1 has the form

$$S_{CE\_net,1} = \sum_j w_j Y_{CE\_net,j} X_{ij} - \left(e_{11} + e_{12}\right).$$

Estimation of the error rates is discussed in Section 7.2.

5. E-sample Errors from the Component Error Perspective

From the component error perspective, enumerations in the wrong location but otherwise correct are correct enumerations, and enumerations with insufficient information for matching could be correct enumerations. The weighted numbers are based on the E sample and are restricted to data defined cases only. Whether a case has sufficient information for matching and whether it went to followup are known without error. We have the following weighted estimates of males (*i*=1) among the enumerations with insufficient information by erroneous and not erroneous enumeration status from the component error perspective:

- Enumerations with insufficient information for matching that did not go to followup
  - CE_{13} = correct enumeration in the correct location
  - NC_{13} = not correct enumeration in the correct location
    - $NC_{13} = EE_{13} + WL_{13}$
    - EE_{13} = erroneous enumeration
    - WL_{13} = in the wrong location but otherwise correct enumeration
- Enumerations with sufficient information for matching that did go to followup
  - CE_{14} = correct enumeration in the correct location
  - NC_{14} = not correct enumeration in the correct location
    - $NC_{13} = EE_{14} + WL_{14}$
    - EE_{14} = erroneous enumeration
    - WL_{14} = in the wrong location but otherwise correct enumeration

Since component error perspective uses both the sufficient and insufficient information enumerations, estimates of the four types of error rates follow where $NE_{1j} = CE_{1j} + WL_{1j}$:

- Cases with sufficient information for matching that did not go to followup
  - r_{1ne} = rate EE miscoded as NE
  - r_{1en} = rate NE miscoded as EE
- Cases with sufficient information for matching that did go to followup
  - r_{2ne} = rate EE miscoded as NE
  - r_{2en} = rate NE miscoded as EE
- Cases with insufficient information for matching that did not go to followup
  - r_{3ne} = rate EE miscoded as NE
  - r_{3en} = rate NE miscoded as EE
- Cases with insufficient information for matching that did go to followup
  - r_{4ne} = rate EE miscoded as NE
  - r_{4en} = rate NE miscoded as EE.

Then estimated errors in $EE_{comp,1}$ for *i*=1 (Male) by the categories follow:

- Cases with sufficient information for matching that did not go to followup
  $e_{11} = (r_{1ne}\, NE_{11} - r_{1en}\, EE_{11})/(1 - r_{1en} - r_{1ne})$
- Cases with sufficient information for matching that did go to followup
  $e_{12} = (r_{2ne}\, NE_{12} - r_{2en}\, EE_{12})/(1 - r_{2en} - r_{2ne})$
- Cases with insufficient information for matching that did not go to followup
  $e_{13} = (r_{3ne}\, NE_{13} - r_{3en}\, EE_{13})/(1 - r_{3en} - r_{3ne})$
- Cases with insufficient information for matching that did go to followup
  $$e_{14} = \left(r_{4ne} NE_{14} - r_{4en} EE_{14}\right) / \left(1 \; - r_{4en} - r_{4ne}\right)$$

Then, the error-adjusted version of EE_{comp} corresponding to Males (*i* = 1) is equal to

$$EE_{comp,1} - \; \left(e_{11} + e_{12} + e_{13} + e_{14}\right),$$

with e_{ij} as just defined. These e_{ij} terms will need to be estimated for subgroups of interest (sample permitting), but not for subgroups defined by the covariates in the logistic regression models. Estimation of the error rates is discussed in Section 7.2.

6. Data for Error Estimation

In the previous two sections, we developed the general form of the estimators for the error components. Now we examine the evaluation studies that will provide data for estimating the error components to see for which we can estimate means, variances, and covariances.

For each of the nonsampling errors in Sections 3, 4, and 5, we will see whether the evaluation studies provide information for estimating their first two moments. In particular, we will assess whether some error components or parts of some error components will not be estimable (or to put things more precisely, whether the moments of some of the error components will not be estimable). In such cases, we will attempt to develop methods of bounding the moments or otherwise quantifying them. We will also look for overlap of evaluation studies, which may provide an opportunity to get better information about the strengths and weaknesses of the studies.

6.1. Further Study of CCM Missed Housing Units (FS)

The FS will analyze data collected and processed for the Evaluation of Address Frame Accuracy and Quality to assess the effect of two types of errors in the census address list on the 2010 Census Coverage Measurement (CCM) estimates of census coverage error. In particular, the study will examine the potential for census geocoding errors and errors in the creation of the initial census address list to introduce bias in the estimates of census coverage error (Mulry, Moran, and Gbur 2011).

The Evaluation of Address Frame Accuracy and Quality is performing an extended search for housing units in the CCM that were coded as "missed" by the P-sample. This study defines a geocoding error as one that places the address more than (alternatively) 1, 3, or 5 kilometers from the correct address. In contrast, the Further Study of CCM Missed Housing Units (FS) defines a geocoding error from the CCM perspective, so that an address is geocoded correctly if it is within the CCM search area (surrounding ring of blocks). To examine the marginal effect of geocoding error for the P-sample nonmatching HUs that are found outside the search area, the FS will recode them as not in the P-sample and recalculate the DSE. The marginal effect is estimated by the difference from the production DSE. A similar analysis can be carried out when more than just this source of error is considered.

The Evaluation of Address Frame Accuracy and Quality (Johnson, 2010) also will determine whether some of the nonmatching P-sample housing units match a HU on the Census Bureau's Master Address File but did not meet the criteria to be downloaded to the Decennial Master Address File. To estimate their marginal effect on the DSE, we can add them to the E-sample as matches and recode them in the P-sample as matches. As before, the marginal effect is estimated by the difference from the production DSE, and a similar analysis can be carried out when more than just this source of error is considered. The estimates of moments will be subject to sampling error, because the study is based on the P sample.

6.2. Matching Error Study (MES)

Matching error may be larger for the 2010 CCM than for previous censuses because the matching operation is now more complicated. The operation now includes processing the results of the computerized search of the entire census rather than just the search area of the sample block cluster. Also, the matches will be attempted on a wider set of cases, including enumerations with insufficient information. To reduce cost, the MES will utilize results of the 2010 CCM person matching quality control operation (Griffin and Mulry 2011) rather than carry out an independent re-match of a subsample of CCM block clusters.

There are four phases in the quality control (QC): clerical geo-coding, clerical residence status coding, before follow-up clerical matching of person, after follow-up clerical matching of person. The QC process is a dependent re-match and correction on samples of records, and error rates are estimated for each phase. (1) Even if 100% of the records in the samples were re-matched and all detected errors were corrected, the non-sample records could still have errors. (2) Furthermore, even the sample records that were not found to have errors could nonetheless have errors, due to dependency. Regarding point (1), the samples were selected probabilistically with a known sampling design so the error rates can still be estimated and applied to the full set of records, nonsampled as well as sampled. Regarding point (2), Griffin and Mulry (2011) plan to use sensitivity analyses to investigate whether there is a correlation between potentially undetected matching error and estimates of component and net error rates.

The MES will be useful for this study because the MES will provide estimates of the estimated matching error rate after the QC operation for each phase of QC, and it will also provide estimates of the effect on net error estimates and on component error estimates of the remaining matching error after each phase of QC. The phase by phase estimates are important for future census planning and operational evaluation. The estimates after all the phases of QC are key for estimating the nonsampling error in estimates of component error.

The estimates will be provided separately for different types of matching error. This is important for estimation of the fine-level error components in the error decompositions provided in Spencer (2009). On an intuitive level, for example, geocoding errors for an in-mover address will have different effects (on net and component error estimates) than errors in coding of residence status. Griffin and Mulry (2011) note that “Ranges of errors and estimates due to the probability sampling (sampling error) and due to errors in QC verification and correction (using the sensitivity analysis) will be produced.” They also note that “Errors remaining in match codes for census and Personal Interview (PI) persons who were coded by technicians during before follow-up processing will be analyzed separately from errors remaining for persons with unresolved residence or enumeration status sent to CCM Person Follow-up.”

6.3. Respondent Debriefings (RD)

The Respondent Debriefings (RD) were designed to address the data collection error that may be a source for errors in identifying usual residence and housing unit population membership in the CCM PI and PFU (Nichols and Childs 2010).

The RDs collected information about the error that occurs between the respondent and interviewer regarding the roster of residents, alternative addresses, and moves. In addition, a PFU form review is based on tape-recorded cases. The two respondent debriefing evaluations included sending experts in residence rules and CCM procedures out into the field to accompany PI and PFU interviewers. These experts audiotaped cases they observed in the field as well as those they debriefed. The taped PI cases are assisting in the analysis, and the taped PFU cases are assisting in the respondent debriefing analysis and the form review. The cases were not selected at random, but rather were selected purposively in a manner that focused on areas with high mobility.

The Respondent Debriefings provide “true measurements” by means of which response errors can be observed for P-sample persons in households where Respondent Debriefings were allowed to take place. “Allowed to take place” means that an Expert accompanied the field interviewer during the PI or PFU. If the set of households where a Respondent Debriefing were allowed to take place were chosen by probability sampling, then approximately unbiased estimates of measurement-error corrected totals and rates could be obtained. There are several problems, however. The number of households were Respondent Debriefings were allowed to take place is small (Nichols 2010, 6). The number of interviewers for whom the Respondent Debriefings were conducted is small, and the number of experts is small. If random sampling were used to choose the households and the assignment of interviewers and experts could be regarded as random, the sampling error in estimates of totals and rates would likely be large. However, since the set of households was not chosen by probability sampling, approximately unbiased estimates will not be available. However, the resulting estimates may be informative and provide guidance about errors.

6.4 Recall Bias Panel Study (RBS)

The RBS focuses on errors in the reporting of mover status caused by respondent recall error (Linse and Pape 2010). 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 RBS is a telephone survey that administers a questionnaire that has questions as similar as possible to the questions about a person’s residence in the PI and PFU. The RBS has a dual frame design with two types samples, one uses random digit dialing (RDD) sampling of landline and cellular telephone numbers and the other selects a sample of address on the Master Address File that match addresses in the April 2010 National Change of Address (NCOA) file from the U.S. Postal Service. There are four panels, the first is a sample of 10,000 RDD housing units while the other three panels are composed of 5,500 RDD housing units and 4,500 movers identified by the NCOA file. The panels are interviewed at different time intervals after April, 2010(Linse and Pape 2011).

The RBS will produce two types of analysis. One estimates the proportion of people reporting that they moved after April 1 from each panel. The other uses the sample of movers identified by the April 2010 NCOA to estimate the proportion of people who correctly report moving in April.

Once the data have been analyzed, comparisons can be made between the RBS respondents’ answers at different times points, and they can also be assessed for the sample members known to be movers. In particular, from the mover frame it will be possible to estimate the conditional probability of a mover misreporting as a nonmover, say $p(NM \mid M)$, and it will be possible to estimate the response error distribution for date of move for movers who did report moving, where only recall bias is considered as an error source. The study design does not yield simple estimates of the conditional probability of reporting being a mover given that one is not a mover, say $p(M \mid NM)$, where only recall bias is considered as an error source. Such an error could arise if a person moved shortly before Census Day and got confused about the date when responding to the interview. The expected error in the observed mover rate is

$$r_e = [\,p(M \mid NM) \times N_{NM} - p(NM \mid M) \times N_M\,] / [M + NM],$$

where N_M and N_{NM} are the numbers of movers and nonmovers in the population. Differences in the observed fraction of movers at 4 different time points can be observed from the RDD sample, and under some assumptions it may be possible to infer $p(M \mid NM)$. Such estimates may be subject to considerable error, however.

It is unclear how the RBS will take into account recall bias from proxy respondents, nor whether proxy responses will be used in the RBS itself. If there is no proxy reporting in the RBS, then the conditional probabilities $p(NM \mid M)$ might be understated.

If there is no recall bias, the sample proportions of movers should not vary with the date of the panel (reporting lag) or the date of the move. Unfortunately, the converse is not true. It is possible for there to be no change in the proportion across panels during March and April but to still have misreporting errors, so long as the errors netted out.

It should be noted that RBS approach does not apply to known non-movers who report that they did move, which could be an issue for people who moved shortly before Census Day and misreported mover status in the CCM. The RBS data will permit estimation of misclassification rates for movers. The RBS data are sample based, and the estimates of error rates will be subject to sampling errors which will need to be estimated.

6.5. Comparison of Census History Study (CCH)

The Comparison of Census History and CCM results will provide some information about geocoding error and possibly other types of E sample errors (Ikeda 2010). Files from several steps of census operations for the selected areas will be merged into a single file, along with the portion of the CCM E-sample and P-sample persons and HU files that cover those areas. The intended result is files containing a history of each person and HU from its first appearance along with the CCM data for that person or housing unit. Combining all the necessary census files into a file suitable for such a comparison for a localized area and comparing the results to the CCM will be a start to improving the evaluation of census operations using data from CCM.

The sample for the CCH will be a subsample of the CCM sample. Since the methodology for the CCH is new, we will not know whether the results will in fact be useful in our study until the CCH is underway.

7. Estimating of Error Components

In this section, we focus on estimating nonsampling errors for matches, correct enumerations from the net error perspective, and the component erroneous enumerations. The component omissions estimate is derived from these three.

7.1. Matches

The discussion in Section 3 shows that it is necessary to to estimate r_{imn} and r_{inm} for subgroups corresponding to all values of all predictor variables X_i. Both field error and processing error contribute to r_{imn} and r_{inm}.

Processing error for each of the six categories may be estimated from the Matching Error Study (MES), which is based on a probability sample of records. The errors as estimated from the MES may understate the actual error, as they are based on dependent rematches. However, if analyses suggest that the rates are too understated, and adjustments are suggested, then they could be incorporated.

Data collection error for each of the six categories may be estimated in part from the Respondent Debriefings (RD), which are administered to small nonprobability samples of the CCM Person Interviews (PI) and CCM Person Followup (PFU) Interviews. The respondent debriefings yield alternative measures of "true" residency status, and by utilizing these rather than the original responses the contributions of field error to the error rates may be assessed.

The RD may fail to detect error in recall, however. The Recall Bias Study is based on probability samples of people independent of the CCM, and it will lead to estimates of underreporting and overreporting of moves in the PI and PFU; misreporting is a concern given that the PI occurs more than 5 months after Census Day and the PFU occurs more than 9 months after Census Day.

Since the RBS interviewed sample in September 2010, which coincides with the timing for the CCM PI, and February 2011, which coincides with the CCM PFU, error rates will be available by mover status at the time of the PI and PFU interviews. However, the RBS will not provide error rates by match status to adjust estimates for nonmovers, movers, and outmovers, so some assumptions will need to be made to distribute the errors to the four components within the nonmovers, inmovers, and outmovers. Other assumptions will be needed to account for the fact that RBS yields information for a general population rather than specifically for persons whose cases would have gone to Followup or not, or those who would match or not.

7.2. Correct and Erroneous Enumerations

For the net error perspective, the discussion in Section 4 shows that it is necessary to estimate r_{icn} and r_{inc} for subgroups corresponding to all values of all predictor variables X_i. These will suffice to estimate $e_{11}$and e_{12} for correcting the sufficient statistics for fitting the logistic regression model for the probability of a correct enumeration that is used in estimating the *DSE* and thereby, the component error estimate of omissions. Both field error and processing error contribute to r_{icn} and r_{inc}.

For the component error perspective, the discussion in Section 5 shows that it is necessary to estimate r_{ien} and r_{ine} for subgroups corresponding to all values of all predictor variables X_i. These will suffice to estimate e_{11}, $e_{12,}$ e_{13}, and e_{14} for correcting component error estimate of erroneous enumerations and thereby, the component error estimate of omissions. Both field error and processing error contribute to r_{ien} and r_{ine}.

Processing error rates may be estimated from the Matching Error Study (MES). As discussed in Section 6.2, the MES is based on a probability sample of records, but the MES uses dependent rematches to assess error, which may lead to understatement. Adjustments for understatement could be utilized if available, however.

Field error rates may be estimated in part from the Respondent Debriefings (RD), which are administered to small non-probability samples of the CCM Person Interviews (PI) and CCM Person Followup (PFU) Interviews. The respondent debriefings yield alternative measures of "true" residency status, and by utilizing these rather than the original responses the contributions of field error to the error rates may be assessed.

The RD will fail to detect misclassification of correct and erroneous enumerations as caused by geocoding error. Information about geocoding error in the P sample will be provided by FS and, possibly, by the Comparison of Census Operations History Study.

In addition, the RD may fail to detect error in recall. However, the RBS is based on probability samples of people independent of the CCM, and it will lead to estimates of underreporting and overreporting of moves in the PI and PFU; misreporting is a concern given that the PI occurs more than 5 months after Census Day and the PFU occurs more than 9 months after Census Day. If a mover is misreported as a nonmover, then an enumeration of the person in the sample block cluster has a high probability of being coded a CE.

8. Conclusion and summary

We have presented a general form for estimators to account for nonsampling errors through adjusting the sufficient statistics for logistic regression estimation. This has been in the context of using logistic regression in dual system estimation as planned for the CCM. In addition, we have examined the data that will be available from evaluations of CCM and assessed its potential for aiding in the estimation of nonsampling errors.

Future work will assess the limitations of the estimators both in terms of their bias and in terms of their variance, which will be important for estimating the moments of the component errors for subgroups. Also, future work will involve modifying the estimators to make them less susceptible to the limitations of the evaluation data for estimating particular nonsampling errors.

The research will seek to use sensitivity analyses to investigate the effects of nonsampling errors on the estimates of net census coverage erroneous enumeration, and omissions. In particular, the research will assess the the combined effect of all the nonsampling errors and identify which sources and types of error have the largest effects. The goal is to provide information for improving the designs of future censuses and census coverage measurement programs.

References

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

Cantwell, P. and Ramos, M. (2010) "Recommendation for the Geographic Level of Estimates Released from the 2010 Census Coverage Measurement Program." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #A-46. U.S. Census Bureau. Washington, DC.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. London: Chapman and Hall.

Griffin, R. and Mulry, M. H. (2011) "2010 Census Program for Evaluations and Experiments Study Plan: 2010 Census Coverage Measurement Matching Error Study." draft 2/10/11. Decennial Statistical Studies Division. U.S. Census Bureau, Washington, DC.

Ikeda, M. (2010) "Census Program for Evaluations and Experiments Study Plan: Explaining How Census Errors Occur through Comparing Census Operations History with Census Coverage Measurement (CCM) Results." Center for Statistical Research and Methdology. U.S. Census Bureau, Washington, DC.

Johnson, N. (2010) 2010 "Census Program for Evaluations and Experiments Study Plan: Evaluation of Address Frame Accuracy and Quality." DSSD 2010 DECENNIAL CENSUS MEMORANDUM SERIES #O-A-3. Decennial Statistical Studies Division. U.S. Census Bureau, Washington, DC.

Katz, J. N., and Katz, G. (2010) "Correcting for survey misreports using auxiliary information with an application to estimating turnout". *American Journal of Political Science*, 54, 815-835.

Linse, K. and Pape, T. (2011) "2010 Census Program for Evaluations and Experiments Study Plan: 2010 Census Coverage Measurement Recall Bias Panel Study Plan." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-I-10R. Decennial Statistical Studies Division. U.S. Census Bureau, Washington, DC.

Linse, K. and Pape, T. (2010) "2010 Census Coverage Measurement Recall Bias Study Plan."DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-I-05. Decennial Statistical Studies Division. U.S. Census Bureau, Washington, DC.

Lyles, R. H., Tang, L., Superak, H. M., King, C. C., Celentano, D. D., Lo, Y., and Sobel, J. D. (2011) "Validation data-based adjustments for outcome misclassification in logistic regression: an illustration." *Epidemiology*, 22, 589-598.

Magder, L. S. and Hughes, J. P. (1997) "Logistic regression when the outcome is measured with uncertainty". *American Journal of Epidemiology*, 146, 195-203.

Morrissey, M. and Spiegelman, D. (1999) "Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons." *Biometrics*, 55, 338-344.

Mulry, M. H., Moran, M., and Gbur, P. (2011) "2010 Census Program for Evaluations and Experiments Study Plan: Evaluation to Assess Effect of Errors in the Census Address List on Census Coverage Measurement Estimates." Draft June 24, 2011. Decennial Statistical Studies Division. U.S. Census Bureau, Washington, DC.

Mulry, M. H. and Spencer, B. D. (2010a) "Developing an Error Structure in Components of Census Coverage Error." *Proceedings of the Survey Research Methods Section,* Alexandria, VA.: American Statistical Association.

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

Nichols, E. (2010) "2010 Census Program for Evaluations and Experiments Study Plan: Respondent Debriefings of the 2010 Census Coverage Measurement Person Interview (CCM PI) and Person Followup (CCM PFU)." Center for Survey Methodology. U.S. Census Bureau, Washington, DC.

Nichols, E. and Childs, J. H. (2009). "Respondent Debriefings Conducted by Experts: A Technique for Questionnaire Evaluation." *Field Methods* 21, 115-132.

Neuhaus, J. M. (2002) "Analysis of clustered and longitudinal binary data subject to response misclassification." *Biometrics*, 58, 675-683.

Paulino, C. D., Soares, P., Neuhaus, J. (2003) "Binomial regression with misclassification." *Biometrics*, 59, 670-675.

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