UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

October 30, 2008

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-E-22

MEMORANDUM FOR Donna L. Kostanich, Co-Chair
Census Coverage Measurement Operational Integration Team

Magdalena Ramos, Co-Chair
Census Coverage Measurement Operational Integration Team

From: Richard A. Griffin *(signed)*
Chair, Census Coverage Measurement Estimation Subgroup

Prepared by: Richard A. Griffin
Decennial Statistical Studies Division

Subject: Methodology Issues

The attached document on "Methodology Issues" is being provided as background material for the Census Coverage Measurement Estimation Workshop to be held in January 2009.

Attachment

cc:
DSSD CCM Contacts List

Methodology Issues

I. Introduction

The 2010 Census Coverage Measurement Survey (CCM) is different than previous census coverage measurement surveys in many ways. Some of the major differences for estimation are the emphasis on the estimation of census component errors and the use of logistic regression modeling in dual system estimation for net error. Eight reference papers have been prepared for this Estimation Review Workshop as background. Mule (2008A) provides a detailed overview of all planned estimation methodology. We are interested in input from the Estimation Review Workshop on any methodology documented in this overview. There are some issues that we have identified as ones we feel warrant more detailed discussion than that provided in the overview. For some of these estimation issues we have prepared a reference paper providing more details. We are especially soliciting workshop input on issues related to these topics. Olson (2008) discusses correlation bias adjustment by individual years of age. Griffin and Mule (2008) discuss handling of Spurious Events particularly duplicate enumerations. Mule and Malec (2008) discuss handling of missing data for component error insufficient information cases. Additionally, the proposed coverage estimates are provided in Mule (2008B). A high level overview of the 2010 CCM survey is given in Whitford (2008) and more details concerning the 2010 CCM design can be found in Moldoff (2008).

Included here are some additional estimation topics for which we would like to obtain reaction from this workshop. These are all important and highlighted here with only minimal background. This document also presents general questions on these additional estimation topics.

II. Additional Topics

Each of these topics correspond to a session on the CCM Review Workshop Agenda. The corresponding agenda topic is given below for each of these additional topics as well as a reference to the section of the estimation methodology overview (Mule (2008A)) that discusses the topic.

A. Model Selection (Net Error; January 12, 1:00 PM)

Do you have any input for deciding on a final set of variables and interactions to use for our net error dual system estimation logistic regression models?

Section 2.1.5 of Mule (2008A) discusses plans for dual system estimation using logistic regression. Since the completion of the Accuracy and Coverage Evaluation Revision II we have done considerable research on potential independent variables to predict match probability, enumeration probability and data-defined probability. We plan on looking at statistics such as the Logarithmic Penalty Function to aid in selecting the independent variables for modeling. In addition we plan on making use of summary statistics and the cross-validation procedure available from SAS PROCLOGISTIC. Partial residual methods suggested by Landwehr (1984) to help identify the functional form for any possible continuous independent variables as well as methods suggested by Roberts, Rao, and Kumar

(1987) to detect influential observations and sensitivity of estimation results, will also be used to aid model selection. Population estimates for domains of interest for 2010 will be computed from research files to evaluate differences between potential models for key estimates. We will also compare the predicted probability from logistic regression models with results from a couple of reasonable post-stratification approaches. Differences could identify possible misspecification or models that may need additional interactions. We expect to have several competing models all of which seem reasonable. We plan on using the same main effects for the data-defined probability, correct enumeration probability, and match probability logistic regression net error models. The missing data logistic regression models should also include at least the main effects used for net error.

B. Household Composition Variables (Net Error; January 12, 1:00 PM)

What are your thoughts or concerns with using household composition variables in our net error logistic regression models?

See, Section 2.1.5 of Mule (2008A)

Current research as well as research prior to Census 2000 have shown that variables that depend heavily on who is captured such as household size or a household composition variable based on household member relationships are powerful predictors of census capture. However, we have not used these variables due to concerns about inconsistency between classifications in the E sample and P sample for these variables. We have attempted to quantify the trade-offs of classification bias with heterogeneity bias and it seems clear that for total population estimates or for estimates for large demographic groups defined by race/origin and tenure, using these variables to reduce heterogeneity is beneficial in spite of this inconsistency. However, estimates for smaller geographic groups or groups defined by inconsistent variables such as married or not married are a concern.

C. Estimating the Data-Defined Component for Net Error (Net Error; January 12, 1:00 PM)

What are your thoughts on how we plan to account for data-defined census records in estimating net error?

See Section 2.1.5 of Mule (2008A)

In previous coverage measurement surveys we have used post-stratification and the calculation of the dual system estimate (DSE) in each post-strata. For each post-stratum the DSE was defined as follows:

$$DSE = DD\left(\frac{C\hat{E}}{\hat{E}}\right)\left(\frac{\hat{M}}{\hat{P}}\right)$$

where

DD = the number of census data-defined persons in the post-stratum

$\hat{E}$ = the weighted E-sample population in the post-stratum

$\hat{P}$ = the weighted P-sample population in the post-stratum

$C\hat{E}$ = the weighted E-sample estimate of correct enumerations in the post-stratum

$\hat{M}$ = the weighted P-sample estimate of matches to the census in the post-stratum

Notice that DSE is the product of three factors: (1) the number of data-defined persons, (2) the estimated proportion of data-defined records that are correct enumerations , and (3) the estimated proportion of persons in the independent P sample who match to the census. The data-defined component is an actual count of the number of data-defined persons in the post-stratum and not an estimated rate.

To calculate synthetic estimates, a coverage correction factor (CCF) was estimated for each post-stratum. The formula for the CCF for a post-stratum is as follows:

$$CCF = \frac{DSE}{CEN} = \left(\frac{DD}{CEN}\right)\left(\frac{C\hat{E}}{\hat{E}}\right)\left(\frac{\hat{M}}{\hat{P}}\right)$$

where CEN is the census count in the post-stratum.

To calculate a synthetic estimate for an estimation domain D, let j indicate the post-strata contained in j. The synthetic estimation for domain D has been calculated by the following formula:

$$SYNDSE_D = \sum_j CEN_{Dj} CCF_j = \sum_j CEN_{Dj} \left(\frac{DD_j}{CEN_j}\right)\left(\frac{C\hat{E}_j}{\hat{E}_j}\right)\left(\frac{\hat{M}_j}{\hat{P}_j}\right)$$

Notice that the synthetic estimate sums over each post-stratum, the census count in that post-stratum and domain multiplied by three factors calculated at the post-stratum level. These factors are respectively a post-stratum level data-defined rate, estimated correct enumeration rate, and estimated inverse of the match rate. Here the data-defined component is the overall post-stratum rate that is assumed to apply for all census persons in that post-stratum and in domain D. Thus, even if an estimation domain has a relatively small number of data-defined cases, the synthetic estimate uses data-defined rates at the overall post-stratum level applied to census counts at the domain level. Since the domain level census counts include both

data-defined and nondata-defined persons, the synthetic estimate is appropriate even with a relatively small number of data-defined cases in the domain.

For 2010, we plan on using the N_0 estimator defined in Section 2.1.5 of Mule (2008A). This estimator sums over all census records in a domain. For each census record in the domain the three logistic regression models are used to predict a data-defined probability, correct enumeration probability and match probability using information on the independent variables for each census record. The product of the data-defined probability and the correct enumeration probability is divided by the match probability. The result is summed over all census records in the estimation domain.

Another estimator, which we call N_1 (See Griffin (2005)), uses a data-defined indicator for all census records instead of a predicted data-defined probability from a logistic regression. Data-defined status is available for all census records so it may seem not necessary or even desirable to use a model. However, the N_1 estimator has a potential weakness for any small estimation domain that has a relatively small number of data-defined census records. Without adding additional variables to the match rate model to compensate for a small number of data-defined cases, the N_1 estimator would likely be too low.

D. Person Omissions by Housing Unit Status (Component Error; January 12, 3:15 PM)

Do you have any ideas on how we might enhance our current plans for estimating person omissions by housing unit status?

For people in housing units, the CCM program has the objective of estimating omissions by whether the housing unit was included or not. It is challenged to estimate one quantity being missed by whether another quantity was missed. There is no simple weighted tabulation available so several assumptions will be necessary. Three possible approaches we are not planning to use are discussed in Section 2.2.5 of Mule (2008A). We have decided to use the same methodology that was used for the 1990 Post-Enumeration Survey. Estimates will be generated by using only the resolved data from housing units that had people who did not move in or out since Census Day. The estimate of person omissions by housing unit inclusion status uses the proportion of resolved nonmover person nonmatches that were in housing units that a) matched or b) did not match (details in Section 2.2.5 of Mule (2008A)).

E. Estimation for Unresolved Enumeration Status for Housing Unit Geocoding Errors (Missing Data for Component Error, January 13, 1:00 PM)

Do you have any ideas on approaches we could investigate so that our imputation for unresolved enumeration status for housing unit geocoding errors could better reflect geocoding error?

See Mule (2008A), section 3.3.5.

The definition of erroneous enumerations for component error is different than the definition for net error. One example of the difference is that an enumeration may be erroneous for net error estimation because it was a geocoding error. For component estimation, a housing unit is correct if it was included once and only once in the census. These geocoding errors have been erroneously assigned to the sample block cluster by the census. For some cases that are unresolved as to whether they are correct or erroneous, we will not be able to tell if they are geocoding errors or do not represent a housing unit. The reason is that our search for them during fieldwork and matching is limited to the block cluster search area. We will not be able to detect a housing unit located outside the search area but geocoded to the sample block cluster. The current plan is to use cell adjustment, where for each defined cell the proportion of resolved housing units that are correct enumerations is imputed as a correct enumeration probability for unresolved housing units in the cell. This is effectively making an ignorable missing data assumption. Defining a cell to only include geocoding error records would help. However, all resolved housing units will be based on information attainable in the block cluster search area. The proportion of these resolved units that are correct may not be representative of housing units where the correct location is outside the block cluster search area. We will have no data available on the enumeration status of housing units where the true location is outside the search area.

References

Griffin, R. (2005), "Net Error Estimation for the 2010 Census", DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-E01.

Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. (1984), "Graphical Methods for Assessing Logistic Regression Models", Journal of the American Statistical Association, Volume 79, No. 385, 61-71.

Roberts, G., Rao, J.N.K., and Kumar, S. (1987), "Logistic Regression Analysis of Sample Survey Data", Biometrika, Volume 74, 1-12.

References provided as background for the Estimation Review Workshop

Griffin R. and Mule T. (2008), "Spurious Events in Dual System Estimation", DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-E-20.

Mule, T. (2008A), "2010 Census Coverage Measurement Estimation Methodology", DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-E-18.

Mule, T. (2008B), "Proposed Census Coverage Measurement Estimates for Net and Component Error", DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-A23.

Mule, T. and Malec D. (2008), "Missing Data Estimation for Component Error Insufficient Information Cases", DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-E-21.

Olson, D. (2008), "Correlation Bias Adjustment by Single Years of Age", DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-E-19.

Moldoff, M. (2008), "The Design of the Coverage Measurement Program for the 2010 Census", DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-B-07.

Whitford, D. (2008), "Overview of 2010 Census Coverage Measurement Program", DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-A19.