Direct Estimates As a Diagnostic for Dual System Estimators Based on Logistic Regression

Mary H. Mulry, Bruce D. Spencer, Tom Mule, Nganha Nguyen[1]
Statistical Research Division, U.S. Census Bureau, Washington, DC 20233
Department of Statistics and Institute for Policy Research, Northwestern University, Evanston, Illinois 60208
Decennial Statistical Studies Division, U.S. Census Bureau, Washington, DC 20233
Decennial Statistical Studies Division, U.S. Census Bureau, Washington, DC 20233

Abstract

The 2010 Census Coverage Measurement Program (CCM) is preparing to use logistic regression modeling in the estimation of net census coverage error rather than poststratification, the approach used for previous censuses. The most important objective for the CCM is to obtain separate estimates of erroneous census inclusions and census omissions. The plan for estimating census omissions is to sum estimates of net coverage error and erroneous enumerations. The net error estimates will be based on dual system estimation formed with separate logistic regression models for the correct enumeration rate and the match rate. Direct estimates at the block cluster level aid in variable selection by comparing the accuracy of estimates based on logistic regression models (or poststratification designs) with and without a variable for groups of the clusters with different characteristics.

Keywords: census coverage error, Accuracy and Coverage Evaluation Survey, 2010 Census Coverage Measurement Program

1. Introduction

The 2010 Census Coverage Measurement Program (CCM) is preparing to use logistic regression modeling in the estimation of net census coverage error rather than poststratification, the approach used for the evaluations of the three previous censuses. The most important objective for the CCM is to obtain separate estimates of erroneous census inclusions and census omissions. The plan for estimating census omissions is to sum estimates of net coverage error and erroneous enumerations.

For the 2010 CCM, the estimates of net census coverage error will be based on models that provide indirect estimates for areas and groups below the national level. The choice of models, whether based on logistic regression or poststratification or other methodology, will require variable selection and other aspects of model selection. Direct estimates of net coverage error can be obtained at the block cluster level, as block clusters are sampling units in the survey for CCM, which is a post enumeration survey. The direct estimates can be used as benchmarks for comparison with indirect estimates produced by alternative models.

In a previous study using such an approach, Mulry, Schindler, Mule, Nguyen, and Spencer (2005) found that comparisons between indirect estimates and direct estimates as well as comparisons between census counts and direct estimates showed large discrepancies, e.g., root mean squared deviations of around 20% and mean absolute discrepancies of around 10% for block clusters with Census 2000 counts of 100 or more. Qualitatively similar results were found by Spencer and Hill (2001) for the 1990 census. One concern in using direct estimates as standards of comparison is that the direct estimates themselves are subject to error. For example, in the studies mentioned above, the levels of discrepancy varied with the choice of direct estimator, e.g., Census-Plus, direct DSE, direct DSE modified by a correlation bias adjustment. Aside from modeling issues in the direct estimates, random errors may arise as well as biases such as geocoding errors.

[1] This report is released to inform interested parties of research and to encourage discussion. The views expressed are the authors' and not necessarily those of the U.S. Census Bureau.

The diagnostic based on direct estimates for large block clusters offers an aggregate assessment that compliments diagnostics for individual logistic regression models. The 2010 CCM estimation of net coverage error uses dual system estimation formed with rates estimated using logistic regression. Since the estimates of the rates are used in a ratio, variable selection may be more complicated than for logistic regression models designed to stand-alone.

In this paper, we apply the diagnostic to compare the accuracy of dual system estimates formed using logistic regression models with different variables. The methodology aids in variable selection by comparing the accuracy of estimates based on models with and without a variable for groups of the clusters with different characteristics. In addition, the method has been helpful in evaluating the effectiveness of one form of the logistic regression estimator and in identifying potential over-fitting of a model.

In the next section, we describe the dual system estimator based on poststratification and logistic regression. Section 3 gives a general framework for understanding the effect of such errors on the comparisons. In section 4, we describe some empirical comparisons with data collected in the 2000 Accuracy and Coverage Evaluation Survey (A.C.E.) and sensitivity analyses that can be done in connection with use of direct estimates to validate variable selection and other modeling aspects.

2. Dual System Estimators

A post-enumeration survey that measures census coverage error is composed of two samples, the enumeration sample (E-Sample) and the population sample (P-Sample). The E-Sample is a sample of census enumerations and designed to measure erroneous enumerations. The P-Sample is a sample of the population selected independently of the census and designed to measure census omissions. The members of households interviewed in the P-Sample are matched to the census on a case-by-case basis to determine whether they were enumerated in the census. Both the 2000 A.C.E. and the 1990 Post-Enumeration Survey (Hogan 1992 1993) used dual system estimation to produce estimates of the population size. The A.C.E. Revision II also used dual system estimation (U.S. Census Bureau 2003).

Equation (1) shows the poststratified dual system estimator for poststratum *i*. A.C.E. Revision II used a more complicated version.

$$DSE_i = Cen_i \times r_{DD,i} \times \frac{r_{CE,i}}{r_{M,i}} \qquad (1)$$

where

Cen_i is the census count for the cross-classification of poststratum *i*;

$r_{DD,i}$ is the census data-defined rate for poststratum *i*, which is the percentage of census enumerations that are not whole person imputations.

$r_{CE,i}$ is the correct enumeration rate estimated by the percentage of the enumerations in the E-Sample poststratum *i* that are correct.

$r_{M,i,}$ is the census inclusion rate estimated by the percentage of individuals in the P-Sample poststratum *i* that match a census enumeration, called the match rate.

Estimation for small areas in 2000 and 1990 used the synthetic assumption that the net coverage error rate is constant within the poststratum. To produce estimates for specific areas or population subgroups first coverage correction factors (CCFs) are calculated by dividing the dual system estimates from equation (1) by the corresponding census counts, i.e., $CCF_i = DSE_i/Cen_i$. .

To produce the estimate for any area or population subgroup a, the CCFs are applied synthetically $\sum_i Cen_{a,i} \times CCF_i$ where the summation is over all the *i* poststrata and $Cen_{a,i}$ is the census count in poststratum *i* for area or subgroup *a* (U.S. Census Bureau 2003).

The data-defined rate, the correct enumeration rate, and the match rate may be estimated in more than one way. The logistic regression estimator that the U.S. Census Bureau plans to use for the 2010 Census coverage evaluation is the following (Griffin 2005) :

$$\hat{DSE} = \sum_{s \in Cen} \hat{\pi}_{dd(s)} \frac{\hat{\pi}_{ce(s)}}{\hat{\pi}_{m(s)}} \varphi \qquad (2)$$

where

$\hat{\pi}_{dd(s)}$ = the predicted data-defined rate for census enumeration *s* based on a logistic regression model

$\hat{\pi}_{ce(s)}$ = the predicted correct enumeration rate for census enumeration *s* based on a logistic regression model for the probability of being a correct enumeration using the E-sample

$\hat{\pi}_{m(s)}$ = the predicted match rate for census enumeration *s* based on a logistic regression model for the probability of matching a census enumeration using the P-sample

φ = the correlation bias adjustment factor (for adult males, distinct for a given age-race group)

The estimate for any area or population subgroup *a* is formed by taking the summation over all the census enumerations for area or subgroup *a* $\hat{DSE}_a = \sum_{s \in a} \hat{\pi}_{dd(s)} \frac{\hat{\pi}_{ce(s)}}{\hat{\pi}_{m(s)}} \varphi$. However, the applications of the DSE using logistic regression in this study do not include the correction for correlation bias. 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).

3. General Framework

For a description of the general framework, denote the direct estimate for area i by X_i, the indirect estimate for area i under a set of three models m for the data-defined rate, the correct enumeration rate, and the match rate by Y_{mi}, and the unknown true value for area i by θ_i.

In the comparisons of direct and indirect estimates, the areas will be block clusters or aggregations of block clusters. Denote biases, variances, and covariances as follows.

$$E(X_i \mid \theta_i) = \theta_i + \mu_{Xi} \qquad \mathrm{Var}(X_i \mid \theta_i) = \sigma^2_{Xi}$$

$$E(Y_{mi} \mid \theta_i) = \theta_i + \mu_{mi} \qquad \mathrm{Var}(Y_i \mid \theta_i) = \sigma^2_{mi} \quad \mathrm{cov}(X_i, Y_{mi} \mid \theta_i) = \sigma_{Xmi}$$

To learn about the typical value of the bias μ_{Yi} in the indirect estimate, we consider the average over n areas, $\mu^2_m = \frac{1}{n}\sum_{i=1}^{n} \mu^2_{mi}$. In our case, the areas are block clusters.

Observe that the expected squared discrepancy between the direct and indirect estimate for an area is affected by differences in biases and the variances and covariance of the direct and indirect estimates.

$$E\left((X_i - Y_i)^2 \mid \theta_i\right) = \left(E(X_i \mid \theta_i) - E(Y_i \mid \theta_i)\right)^2 + \mathrm{Var}(X_i - Y_i \mid \theta_i)$$
$$= (\mu_{Xi} - \mu_{mi})^2 + \sigma^2_{Xi} + \sigma^2_{mi} - 2\sigma_{Xmi}.$$

The bias in μ^2_m may be estimated by $\sigma^2 - \bar{\mu}^2_m$ if we assume $\bar{\mu}_m = \bar{\mu}_X$ and the covariances $\sigma_{Xmi} = 0$.

We may obtain a naive estimate of average squared bias μ^2_m by ignoring the biases μ_{Xi} in the direct

estimates. In such a case, we may estimate μ_m^2 by $\hat{\mu}_m^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2 - \hat{V}_{mi}$, where $\hat{V}_{mi}$ denotes the estimates of $\mathrm{Var}(X_i - Y_i \mid \theta_i)$. For details of the derivation, see Spencer (2007).

The bias in the naive estimate can be positive, overstating the average squared bias in the indirect estimates, if there is large variability among the biases in the direct estimates for the small areas. The bias can be negative, on the other hand, if the rel-variance of the biases in the direct estimates $\tilde{\sigma}_X^2 / \bar{\mu}_X^2$, is less than 1. To the extent that the same component of bias is present in the direct estimate and the indirect estimate for an area, the component of bias will not be detected in $\hat{\mu}_m^2$. If such bias components do not covary systematically with model misspecification bias, the estimator $\hat{\mu}_m^2$ will tend to quantify model misspecification bias (including variable selection bias and synthetic estimation bias).

4. Methodology

Direct estimates at the block cluster level formed using the DSE in Equation (1) provide the basis for the diagnostic to assess the accuracy of the dual system estimator, whether based on logistic regression or poststratification. In particular, the method focuses on large blocks, meaning blocks with a census count of at least 100 people, for forming estimates using only data for the block itself. If there was subsampling within the block, the estimates for the block use weighted data and the sampling variance is small. If there was no subsampling, there is no sampling variance. Small or no sampling variance is advantageous is using the direct estimates for large block in a diagnostics. If the ratio of the direct DSE to the Census-Plus estimate[2] for the block cluster was greater than 1.2 or less than 0.8, the direct DSE was capped at the Census-Plus estimate as in Mulry et. al. (2005).

For our study, we index sets of alternative models for indirect estimates by *m* Each alternative model has three models, one for data-defined rate, one for correct enumeration rate, and one for match rate. Also suppose we have sets of block clusters for which direct estimates can be constructed. In our case, these are blocks with a census count of at least 100. The direct estimates for these blocks may use coarse poststratification, for example, to reduce correlation bias in direct DSEs. The direct estimates are formed in a manner consistent with the indirect estimates they use the same data and there is no correction correlation bias. As a test, the sum of the direct estimates for the set of blocks is close to the sum of indirect estimates for the same areas.

Compute the naive estimate $\hat{\mu}_m^2$ for each model m and compare across models. The differences $\hat{\mu}_m^2 - \hat{\mu}_{m'}^2$ for methods m and m' will be subject to sampling variance, in addition to the approximations considered in Section 3. In order to estimate the sampling variance of $\hat{\mu}_m^2 - \hat{\mu}_{m'}^2$ it may be appropriate to consider using replication methods based on random groups.

Formulas for the squared bias estimates for the models are

Unweighted bias: $\hat{\mu}_{mi}^2 = (1/n) * \sum (Yi - Xi)^2 - \hat{V}_{mi}$

Weighted bias based on *Xi:* $\hat{\mu}_{mi}^2 = (1/\sum Xi) * \sum Xi * [(Yi - Xi)^2 - \hat{V}_{mi}]$

Unweighted relative bias: $R\,\hat{\mu}_{mi}^2 = (1/n) * \sum (Yi/Xi - 1)^2 - \hat{V}_{mi}$

Weighted relative based on Xi bias: $R\,\hat{\mu}_{mi}^2 = (1/\sum Xi) * \sum Xi * [(Yi/Xi - 1)^2 - \hat{V}_{mi}]$

[2] Census-Plus = E-sample correct enumerations + P-sample total – P-sample nonmatches.

In initial studies, we discovered that the variance term was very small and did not affect the outcome appreciably. Since calculating was computer-intensive, we decided to assume the variance term was zero.

5. Models

The study examined estimates formed with models using either postratification or logistic regression. The synthetic estimation was based on DSE in Equation (1) and derived with the 416 poststrata used in the A.C.E Revision II. This DSE used the PES-C form of the DSE where the match rate is estimated using the outmovers and the number of movers are estimated using the inmovers as described in Mulry et al (2005). The logistic regression modeling for match status used only nonmovers and outmovers since they were the ones matched in 2000. This results in PES-A estimates of the population. The PES-A national population estimate is about 2 million less than the PES-C estimate. No correlation bias adjustments were used in either indirect estimator.

The basic set of demographic variables that were included both in the logistic regression models and the poststratification model are known as ROAST, standing Race/Origin Hispanic, Age, Sex, and Tenure (owners and renters). The Race/Origin Hispanic variable had seven domains: Non-Hispanic White and Other (intercept), Non-Hispanic Black, Hispanic, Non-Hispanic Asian, Native Hawaiian and Pacific Islander, American Indian on Reservation, and American Indian off Reservations

The use of a spline for Age variable arose from the observation of the match rate and correct enumeration rate had four possible distinct parts: (1) quadratic relationship from 0 to 17, (2) linear relationship from 17 to 20, (3) quadratic relationship from 20 to 50; and (4) linear relationship from 50 on. Using the notation in Smith (1979) , this four part relationship was expressed by the following six covariates in a logistic regression model for match rate. The same model form was used for the correct enumeration rates.

$$Logit(Mrate) = B_0 Int + B_1 \times Age + B_2 \times (Age^2 - (Age - 17)_+^2 + B_3 \times (Age - 17)_+ + B_4 \times (Age - 20)_+$$

$$+ B_5 \times ((age - 20)_+^2 - (Age - 50)_+^2 + B_6 \times (Age - 50)_+$$

The variables in the logistic regression modeling that described the block clusters were Black race rate, Hispanic ethnicity, multi-unit rate, and renter rate. Each rate was transformed for use in the modeling to "arates" by the transformation arate = ln(rate + 1), e.g., proportion Hispanic may be re-expressed as ahisp = ln(hisp +1), where "hisp" refers to the proportion Hispanic.

Other variables that described the census in the block clusters were the type of enumeration area (TEA), whether the block cluster was in a metropolitan statistical area (MSA), and the census mail return rate (cretrate). The cretrate variable was transformed by the same transformation as the other rates. When the transformed variable acretrate appeared in a model, both linear and quadratic terms were used.

A variable for region of the U.S. also was included in some models in attempt to capture any regional variation in the data. This variable had four levels: Midwest, Northeast, South, and West.

Different sets of variables, mostly main effects, were used in fitting a variety of logistic regression models. The models that we discuss in this paper are summarized in Table 1.

Table 1. Logistic regression models considered in comparisons of squared bias

Logistic Regression Model	Variables in model (total number of parameters)
LR	ROAST (15)
LRC	ROAST, cluster rates (19)
LRR	ROAST, region (18)
LRRCX	ROAST, region, cluster rates, MSA, TEA, Mailback rate (29)
LRRCXI	ROAST, region, cluster rates, MSA, TEA, Mailback rate, 2-way interactions (366)

When comparing two models, the squared bias estimates used 2,067 block clusters in the 2000 A.C.E. sample that had a census count greater than or equal to 100. The squared bias was estimated using all the block clusters meeting the size criterion. In addition, estimates of the squared bias were made when the block clusters were divided into groups based on whether their values of different variables were low, medium, or high. The goal of the comparisons of the squared biases for different ranges of values of the variables was to determine whether the inclusion of the variable was improving blocks with extreme values at the expense of the blocks with more moderate values or vice versa. If such trade-offs are to be made, one would be aware of the amount of the trade-off. In addition, block clusters were grouped by the ratio of the capped direct DSE divided by the Census to determine how the model performed in high undercount, high overcount, and more moderate coverage error situations. Table 2 shows the groupings for the continuous variables and the number of block clusters in each group. Comparisons also were done for the region variable where the number of block clusters was 418 in the Midwest, 403 in the Northeast, 629 in the South and 617 in the West.

Table 2. Groupings of 2,067 large block clusters for comparisons of squared biases and number of block clusters in each group

grouping variable	low value	high value	medium value
DSE/Census	< 0.80 n = 89	>= 1.2 n = 27	[0.80 to 1.20) n = 1951
Blackrate	< 0.08 n=1403	<= 0.62 n=203	[0.08 to 0.0620) n=456
Rentrate	<0.51 n = 1405	>=0.86 n=207	[0.51 to 0.86) n = 410
Hisprate	<0.09 n = 1325	>= 0.54 n=257	[0.09 to 0.54) n = 485
Multirate	< 0.55 n = 1452	>= 0.95 n = 216	[0.55 to 0.95) n = 399
Cretrate	< 0.73 n = 1156	>= 0.88 n = 175	[0.73 to 0.88) n = 736

Since we performed 19 tests when comparing each pair of models, we made a Bonferroni adjustment for 19 tests. We looked for differences 3 or more times their estimates standard errors, corresponding to an alpha level for a two-sided test of 0.05 or less. In particular, we noted estimates of difference in squared bias of at least 3.1 times the standard error of the difference.

6. Findings

The comparisons of the squared bias estimates of different models illuminated several different types of problems with variables, model fit, and an estimator.

6.1 Assessing variables

The results of comparisons between the LRR model and the Synthetic model indicated that LRR model is superior to the Synthetic model in terms of unweighted loss, relative loss, and weighted relative loss. Only one group had a statistically significance difference favoring LRR in the unweighted loss comparisons, but the other two loss functions found significant differences for several groups, and all favored LRR.

Comparisons of the squared bias estimates for the LRC and LRR models suggest that they perform about equally well within the resolution of our analyses, but the LRC model is better in terms of relative loss in general, except for a couple of subgroups. These subgroups are DSE/Census >= 1.2 and Cretrate >= 0.88. There is no discernable pattern for weighted relative loss.

The results comparing the LR model to the LRR model showed that the inclusion of region as a predictor variable does not improve accuracy much, and may even decrease the accuracy for some groups. The differences in squared biases using unweighted loss are not statistically significant. The analysis of relative

loss supports the LRR model for the Midwest and Northeast and the LR model for the South. Weighted relative loss comparisons also support the LRR model for the Midwest and the LR model for the South.

6.2 Assessing model fit

The comparison of the LRRCX models with the version that also included the two-way interactions (LRRCXI) found there were no statistically significant differences in unweighted loss. There were some statistically significant differences in relative loss and weighted relative loss favoring the smaller model (LRRCX) over the larger model (LRRCXI) except for the group of bock clusters where the Blackrate was at least 0.62. This suggests that the LRRCXI model is being over-fitted, and that too many interactions are being included in the model. The next step would be to fit models that selectively dropped some interactions to see if the results changed to favor a model with fewer interactions.

6.3 Assessing alternate estimator

The U.S. Census Bureau's initial research for the 2010 CCM also considered another form of the logistic regression estimator. This estimator was known as the *N1* estimator (Griffin 2005), and the U.S. Census Bureau had used this form for its poststratified estimator in 1990 and 2000. The *N1* estimator differs from the estimator in Equation (2) by using an indicator of whether an enumeration was data-defined rather than an estimate of the data-defined rate.

$\hat{N}1 = \sum_{s \in Cen} p_{dd}(s) \frac{\pi_{ce}(s)}{\pi_{m}(s)}$ where $p_{dd}(s) = 1$, if enumeration s is data-defined, and *0*, otherwise.

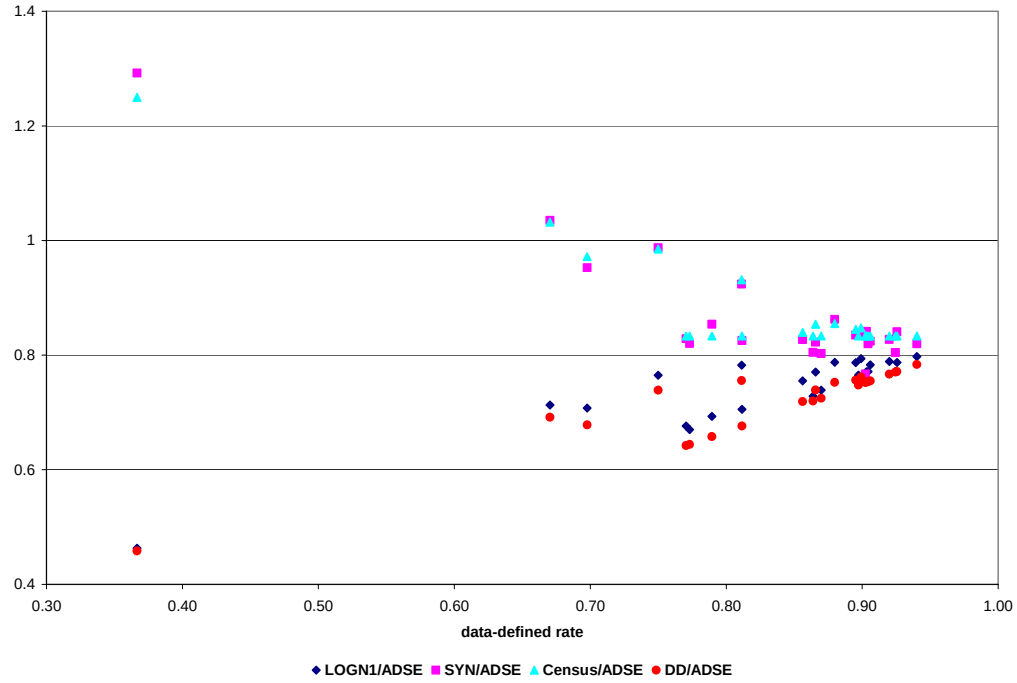
Comparisons of the squared biases for the N1 estimator using the LR model and the Synthetic model produced an interesting result for the Renter groupings of the block clusters. The difference in the squared biases for the middle group had a difference sign than the differences for the high and low groups. Although the differences were not statistically significant, finding further explanations seemed appropriate.

Figure 1 shows plots of the ratio of the *N1* estimate (LOGN1) to the capped direct estimate (ADSE) and of the ratio of the synthetic estimate (SYN) to the capped direct estimate for 25 block clusters where the ratio of the *N1* estimate to the direct estimate was small, below 0.80. Further investigation showed that many of these 25 block clusters had unusually low percentage of persons who were data-defined in the census. The data-defined rates ranged from 37 percent to 94 percent, with an average of 83 percent. Figure 1 also has plots of the ratio of the number of data-defined (DD) to the capped DSE and the census count (Census) to the capped DSE. Dividing by the capped DSE puts all the estimates for the block on the same scale. The Synthetic estimate has the known characteristic of not changing the census count very much as shown in Figure 1. However, the new finding was that the *N1* estimator is very close to the data-defined count when the percentage of data-defined enumerations is low. This type of impact was not present for the estimator in Equation (2) which led to its selection for the estimator to use for 2010.

6. Summary

Direct estimates for large block clusters have proven to be a valuable diagnostic for logistic regression models that are used in dual system estimators. The diagnostic has proven useful in evaluating the aggregate effect of models in addition to individual model assessment for the logistic regression DSE. This diagnostic has aided in the development of models for use in the logistic regression DSE through validating variable selection and assessing models for over-fitting. In addition, the diagnostic has shown that the logistic regression DSE selected for use in the 2010 CCM is more effective than the poststratified DSE and more effective than an alternate form of the logistic regression DSE. Direct estimates for large block clusters will continue to be a valuable asset in the research for the 2010 CCM.

Figure 1. Ratios of Indirect/Direct estimates vs. data-defined rate for 25 blocks



Acknowledgement

The authors thank Eric Schindler for many helpful discussions.

References

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

Griffin, R. A. (2005) "Net Estimation Error for the 2010 Census". DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E01. Decennial Statistical Studies Division. U.S. Census Bureau. Washington, DC.

Hogan, H. (1993) "The 1990 Post-Enumeration Survey: Operations and Results", *Journal of the American Statistical Association*, 88, 1047-1060.

Hogan, H. (1992) "The 1990 Post-Enumeration Survey: An Overview", *The American Statistician*, American Statistical Association, Alexandria, VA. 261-269.

Mulry, M. H., Schindler E., Mule T., Nguyen N., and Spencer B. D. (2005) Investigation of extreme estimates of census coverage error for small areas. *Proceedings of the Survey Research Methods Section*. American Statistical Association. Alexandria, VA.

Smith, P. L. (1979) "Splines as a useful and convenient statistical tool". The American Statistician. 33. American Statistical Association. Alexandria, VA. 57 – 62.

Spencer, B. D. (2007) "Further Investigation of Methodology for Validation of Variable Selection for Net Census Coverage Error Modeling". Report to the U.S. Census Bureau. Spencer Statistics, Inc. under order number YA132306SE0513. August 17, 2007.

Spencer, B. D. and Hill, J. M. (2001) "Activity 3: Final Report on Evaluation of Block Level Estimates". Report to the Bureau of the Census. Abt Associates Inc. and Spencer Statistics, Inc. under contract 50-YABC-7-66020. December 23, 1998; rev. July 12, 2001.

U.S. Census Bureau (2003b) "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