Weight Trimming in the Accuracy and Coverage Evaluation

Vincent Mule
Bureau of the Census, Washington, DC 20233

This paper gives a review of the Accuracy and Coverage Evaluation (A.C.E.) weight trimming procedure. The procedure was designed to protect against undue influence from a small fraction of the sample.

1. Introduction

The A.C.E. weight trimming procedure was designed to reduce the sampling weights for clusters that potentially could have extreme influence on the dual system estimates and variances. The measure of cluster influence was the net cluster error, the absolute difference between the weighted estimate of omissions and the weighted estimate of erroneous enumerations. When the net error exceeded a pre-set maximum value, the sampling weights were reduced. If weights were reduced, this approach reduced variance and may have introduced some bias, but was expected to reduce the mean square error for most estimates. If the net error did not exceed the pre-set maximum value, the sampling weights were unchanged.

After the A.C.E. person matching operation was complete, the net error criteria were examined and, if necessary, weights were adjusted prior to the Missing Data process. If the criteria for weight trimming was met, it was done for all sample cases in a cluster even though a cluster contributed sample to multiple post-strata. For more information on the person matching operation, see Childers 2001. For more information on the Missing Data process, see Ikeda and Cantwell 2001.

The A.C.E. consisted of two independent samples. The first was a sample of the population in the selected A.C.E. sample block clusters, known as the P sample. By matching these people to census records, an estimate of the proportion of the population that was missed in the census was determined. The second was a sample of the census enumerations in the same A.C.E. sample block clusters, known as the E sample. Using the results of matching the P sample to the census, checking for duplication among the census records and re-interviewing to determine correct inclusion of the census record, an estimate of the proportion of erroneously included records in the census was determined. Census 2000 A.C.E. planned to form dual system estimates for post-strata to represent the population in housing units in the 50 states and the District of Columbia (Haines 2000).

It is possible that certain clusters may have exerted undue influence on post-stratum estimates and variances. These are expected to be due to a disproportionate number of census omissions or census erroneous errors within the block cluster. Although extreme sampling weights can be a source of influence in surveys, the A.C.E. sampling weights, the inverse of the probability of selection, were reasonably controlled during survey design and were not expected to be a major contributor to influence for A.C.E.

Thomas Mule is a mathematical statistician in the U. S. Census Bureau's Decennial Statistical Studies Division. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

The weight trimming plan for the 1990 Post-Enumeration Survey (PES) involved trimming the weights of two clusters. Two small clusters with low probability of selection were down-weighted such that the net error of the cluster was roughly equal to the largest net error from the non-small clusters. This was roughly a net error of 150,000. This weight trimming was done for the original 1392 post-stratum design.

The A.C.E. weight trimming plan was a modification of the 1990 method. As in 1990, the cluster weight was trimmed to yield a pre-specified net error. The intention of the plan was to lessen the impact of extremely influential clusters on the dual system estimates and variances.

2. Methodology

Identify Outlier Clusters

Cluster Influence. The measure of cluster influence was the net error. For purposes of weight trimming, the net error was the absolute difference between the weighted number of omissions and the weighted number of erroneous enumerations. The form of the weighted net error is below:

$$Z_i = \left| \left(P_i - M_i\right) - \left(E_i - CE_i\right) \right| \qquad (1)$$

Z_i = the net error estimate for cluster i,
P_i = the weighted P-sample population estimate for cluster i,
M_i = the weighted P-sample match estimate for cluster i,
E_i = the weighted E-sample population estimate for cluster i, and
CE_i = the weighted E-sample correct enumeration estimate for cluster i.

Outlier Criteria. The outlier criterion was the maximum allowable net error for a single cluster. There were three different criteria based on the cluster geography. The nation was classified into two levels of geography: American Indian Reservations and the balance of the nation. The American Indian Reservation clusters were sampled at disproportionately higher rates relative to the balance of the country. In addition, separate American Indian on American Indian Reservation post-stratum estimates were planned. If the American Indian Reservation clusters were included with the rest of the nation, it is unlikely that an influential cluster would be detected. The outlier criteria are defined in Table 1.

Table 1: Outlier Cluster Criteria

Maximum Net Error	Cluster Geography
75,000	Balance of the U.S.
6,250	American Indian Reservation

All clusters with net error greater than the maximum allowable net error were outlier clusters. These clusters were expected to disproportionately influence the dual system estimates and variances, and the sampling weight of the outlier clusters were decreased.

The maximum net error for the balance of the country was based on the 1990 PES. Since the A.C.E. was roughly double the PES sample size, the maximum allowable net error was half the 1990 value. For the American Indian Reservation clusters, the maximum allowable value was a function of the average sampling rates. The American Indian Reservation average P-sample cluster sampling weight was approximately 12 times

less than the Balance of the U.S. average P-sample cluster sampling weight. Because of this, the American Indian Reservation maximum allowable net error was 12 times less than the Balance of the U.S. criteria.

Implementation Strategy. The outlier clusters were identified after the person matching operation was completed, but before the missing data process. The person matching results were the major input into this process. This timing had several implications.

The non-movers and out-movers were used for deriving the estimate of omissions in (1) above. For DSE estimation, if the number of out-movers in a post-stratum was less than 10 then only the non-movers and out-movers were used. Because of the small number of movers expected in most clusters, this process only used non-movers and out-movers.

Some non-movers and out-movers had unresolved match status and residence status. Some E-sample cases had unresolved enumeration status. This meant the status of unresolved cases had to be estimated to identify outlier clusters. Information available at the time of the weight trimming process was used to approximately estimate the unresolved status cases. Since the weight trimming process was done before the Missing Data process, there was some information that the Missing Data process used to estimate unresolved status that was not yet available.

A P-sample non-interview adjustment was estimated in the estimate of omissions. Information available during the weight trimming process was used to approximately estimate the non-interview adjustment for each cluster. Again since the weight trimming process was done before the Missing Data process, there was some information that the Missing Data process used to do the non-interview adjustment that was not yet be available.

The Targeted Extended Search results and sampling rates were reflected in the estimate of omissions and erroneous enumerations.

Because of the implementation issues raised above, the weight trimming estimate was the best estimate of cluster net error at that time that was operationally feasible.

Down-weighting Outlier Clusters

All outlier clusters were down-weighted such that the cluster contributed the maximum allowable number of net errors for the appropriate geography. A separate down-weighting factor was computed for each outlier cluster. The down-weighting factor was the ratio of the outlier cluster criteria to the cluster net error computed in (1).
where

$$D_i = \frac{C}{Z_i} \qquad (2)$$

D_i = the down-weighting factor for cluster i,

C = the maximum net error from Table 1 for the appropriate level of geography, and

Z_i = the net error estimate for cluster i from (1).

The cluster down-weighting factor was applied to the P-sample and the E-sample weights of the outlier clusters. The P-sample and E-sample weights for the remaining clusters were unchanged.

One cluster in the balance of the United States was down-weighted. The estimated net error for the cluster was 77,975. The cluster weights were down-weighted by 0.9618. No AIR clusters required down-weighting.

Figures 1A and 1B show the distribution net error for the two geographic areas prior to weight trimming. These distribution are highly skewed to the right. The down-weighted cluster in the balance of the U.S. was at the extreme tail of the net error distribution.

3. Analysis

Timing

As an evaluation of the weight trimming procedure, the net error estimates were recalculated using the results of the missing data process. The Missing Data results include the non-interview adjustment, imputed residency and match probabilities for the unresolved P-sample cases and imputed correct enumeration probabilities for unresolved E-sample cases.

Using the Missing Data results, revised estimates of net error were calculated. The down-weighted cluster had a revised estimated net error of 77,972. No other clusters were determined to be outliers based on their revised net error estimates. The distributions of net error for both geographic areas were very similar for the original and revised estimates.

The non-interview adjustment and the imputation of unresolved probabilities in the weight trimming estimates provided good estimates of the net error of the cluster.

Change in Mean Square Error

The down-weighted cluster contributed sample to 37 of the 416 post-strata. In these post-strata, we compared the trimmed and untrimmed dual system estimates (DSE_t and DSE_u, respectively) by estimating the relative change in Mean Square Error (MSE) by the following:

$$\frac{MSE(DSE_t) - MSE(DSE_u)}{MSE(DSE_u)}$$

where

$$MSE(DSE_t) - MSE(DSE_u) = Var(DSE_t) + Var(DSE_t - DSE_u) - E\left(DSE_t - DSE_u\right)^2 - Var(DSE_u)$$

and

$$MSE(DSE_u) \cong Var(DSE_u).$$

Figure 2 shows the relative change in MSE for the post-strata. The results show no significant change in mean square error for these 37 post-strata.

The Non-White Hispanic Owner Large Metropolitan Statistical Area Mailout/Mailback South Region High Return Rate 30-49 Male post-stratum had the largest decrease in relative change in MSE. The down-weighted cluster contributed a net error of 7,755 to the estimates. About 99 percent of the clusters contributing to this post-stratum had a net error less than 2,000. Down-weighting this outlying observation slightly lowered MSE for the post-stratum.

The largest increase for a post-stratum occurred because the down-weighted cluster contributed zero net error to the estimate. For this post-stratum, the down-weighting only added weight variation to the estimates.

Result

The weight trimming for the Accuracy and Coverage Evaluation used a good approximation for the Missing Data procedures.

The weight trimming had a minimal impact on the Mean Square Error of the post-strata estimates. No significant changes in Mean Square Error were seen.

4. Future Research

Areas for further research in weight trimming in coverage evaluations include:

- Exploring methodology to focus on implementation at the post-stratum level.

- Explore alternative methodologies using robust estimation methodologies to handle outliers.

References

Childers, D., "Accuracy and Coverage Evaluation: The Design Document," DSSD Census 2000 Procedures and Operations Memorandum Series S-DT-1, U.S. Census Bureau, Washington, D.C., 2001

Haines, D., "Accuracy and Coverage Evaluation Survey" Final Post-stratification Plan for Dual System Estimation," DSSD Census 2000 Procedures and Operations Memorandum Series Q-24, U.S. Census Bureau, Washington, D.C., 2000.

Ikeda, M. and Cantwell, P., "Missing Data Procedures for the Census 2000 Accuracy and Coverage Evaluation Sample," Presented at the American Statistical Association Joint Statistical Meetings, August 6, 2001.

Figure 1a: Distribution of Net Error for the Balance of the U.S.

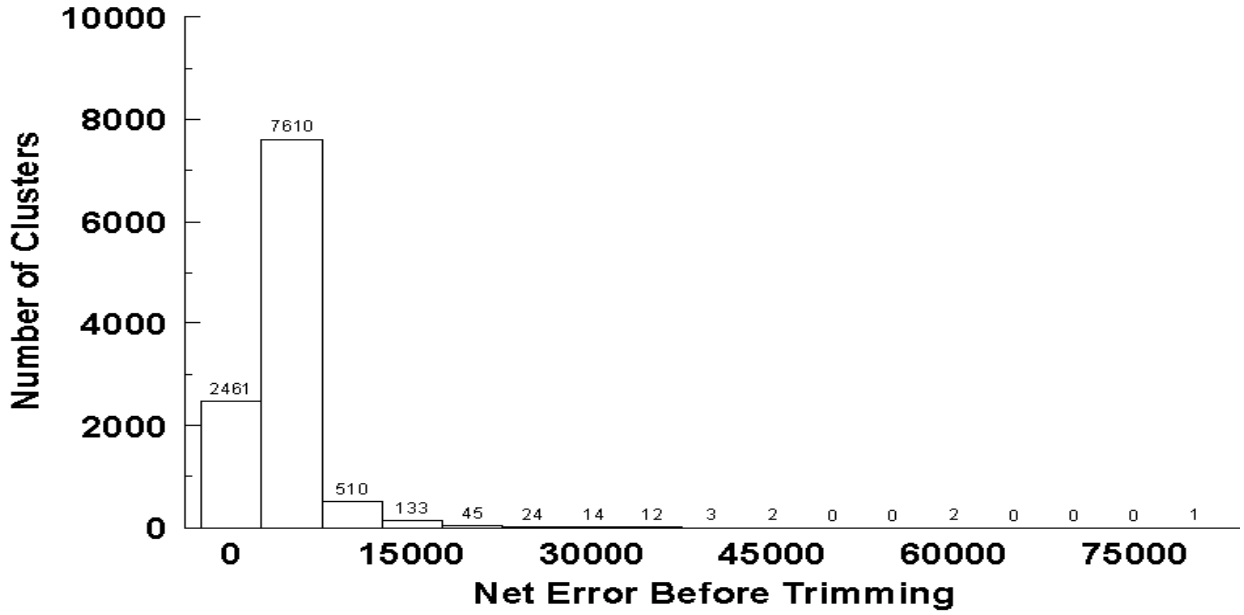

Figure 1b: Distribution of Net Error for AIR

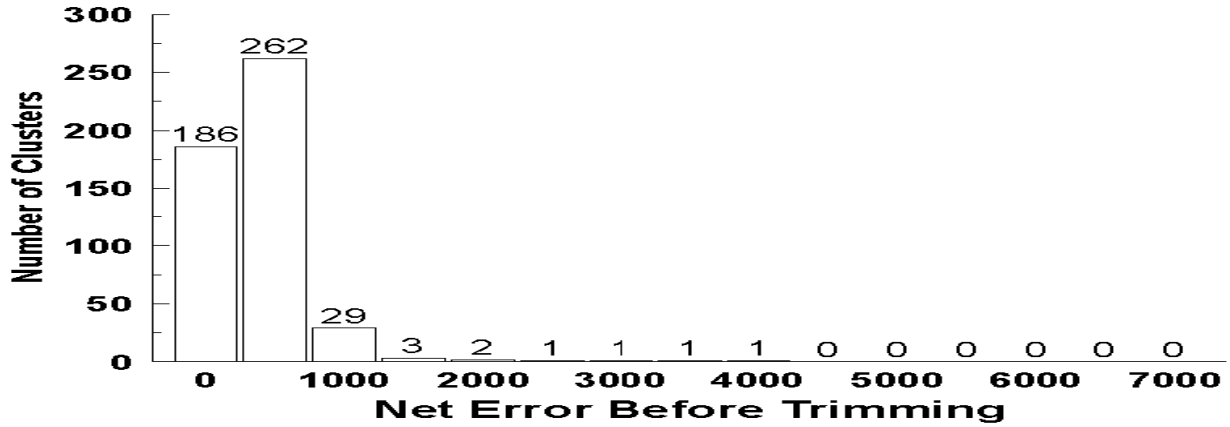

Figure 2 Estimated Change in MSE from Weight Trimming