ACCURACY AND COVERAGE EVALUATION SURVEY TARGETED EXTENDED SEARCH

Alfredo Navarro and Douglas B. Olson, US Census Bureau
Alfredo Navarro, 2118-2 DSSD, Washington DC, 20233-7613

Key Words: Geocode Error, Balancing, Whole Household non-match, Jackknife

I. Background

The 1990 Decennial Census used a Post-Enumeration Survey (PES) methodology to measure census undercount or overcount. The Census Bureau used the dual-system model to produce an estimate of total population. This model relies on classifying each person from the "true" population as being either included in the Census or not, as well as being included in the PES or not. The 1990 PES was conducted on a sample of blocks and used the dual-system estimator (DSE) to produce estimates of total population for various demographic groups. The DSE estimation rules were very simple. For example, a person was considered a census enumeration if he or she was tallied in the population count. This person or census enumeration was considered a correct enumeration if he or she was supposed to be included in the census. On the contrary, a person counted in the census when he or she was not supposed to be included was termed an erroneous enumeration. A person was classified as a census omission if he or she was supposed to be in the census enumeration but was not. There was also a requirement for people to be counted in the "right location."

In the context of the 1990 Census PES, right location meant anywhere in the block cluster where the reported housing unit address was located. A block cluster is a group of blocks with an average size of 30 housing units. In addition, the 1990 PES design employed the concept of search area. For the most part, the search area was defined as one ring of adjacent blocks. In rural areas, the search area was expanded to two rings of surrounding blocks. In List/Enumerate areas, the entire address register area was searched. Using search areas, as long as a census person was counted in the correct block or in any of the blocks in the ring(s) of surrounding blocks it was labeled as a correct enumeration. Persons in the P-sample found in the search area were treated as matches, that is, as not missed by the census. If these two effects are balanced the measure of net undercount is not affected. Failure to balance the two effects results in "balancing error".

The process of searching identifying census omissions and erroneous enumerations is very costly, labor intense, and requires highly trained and skillful clerks (Hogan, 1993.) The resources required to repeat the 1990 PES extended search operation in the 2000 A.C.E. are extremely difficult to effectively operationalize. Limiting the amount of searching to block clusters with a potential high payoff will result in a more accurate and efficient search operation. For the 2000 Census A.C.E. the concept of right location is more limited than the concept used in the 1990 Census PES. The search itself will be targeted or limited to persons in specific households, namely geocoding errors of exclusion and inclusion. Geocoding errors of exclusion affect the P-sample match rate (census coverage rate.) Geocoding errors of inclusion affect the E-sample erroneous enumeration rate. A third major difference is that the 2000 A.C.E. search operation will be sample based. The 1990 PES performed a search for all block clusters in sample.

The PES implementation or the DSE does not require an extended search operation. The expectation of the DSE is not affected by the introduction of the search area concept. In other words, if the search area is limited to the PES block cluster (as in the 1995 Census Test) the expectation of the DSE is the same as that under the 1990 PES search area definition. The motivation for using an extended search area definition is variance reduction. Allowing more cases to be matched and more census enumerations found in surrounding blocks result in a higher match and correct enumeration rates and a DSE with more precision or less variance.

This paper describes the methodology for targeting, sampling, and operational issues associated with the A.C.E extended search plans for Census 2000. The concepts of "targeting" and "balancing error" are further discussed in Section II. In addition, section II describes search plans for 2000. Section III describes the targeting criterion and sampling operations used for identifying and selecting the TES block clusters. It also gives some results from empirical research performed to compare alternative sample designs and estimators considered for implementation in Census 2000. Section IV describes the extended search operations for persons. The final section highlights the effects of the TES on dual system estimation.

II. Search Plans for 2000

The 2000 Census A.C.E. search operation differs from the 1990 PES in three areas, these are: search area definition, amount of searching, and eligible people for searching.

The search area for the 2000 A.C.E. will be limited to either just the sample block cluster or at most one ring of surrounding blocks. A block is in the search area if it touches the cluster of sample blocks at one or more points. This definition includes the blocks that touch the corner of the block cluster. Results from empirical research using Census 2000 Dress Rehearsal data show that the additional benefits of using two rings of surrounding blocks are almost negligible (Wolfgang, 1999.) Additionally, the plan is to implement a sample based search operation. The targeting will be implemented in two phases. Specifically, the plan calls for targeting the extended search operation to 20 percent of the A.C.E. sample blocks. The second phase limits the searching of census omissions and erroneous enumerations due to duplication within the surrounding blocks to the block where the housing unit is found. The search will be concentrated in block clusters thought to have the biggest payoff in terms of variance reduction. Targeted extended search or TES is the manner in which we will determine which block clusters to search. Census geocoding errors affect both the census omission rate (or the P-sample match rate) and the census erroneous enumeration rate. Block clusters with high concentration of census geocoding errors can be identified from the results of the after-followup housing unit clerical matching operation. These are A.C.E. block clusters with a large number of P-sample housing units not found in the E-sample and referred to as whole household non-matches. This type of non-matches are possibly census geocoding errors of exclusion. On the E-sample side, these are A.C.E. block clusters with a large number of census geocoding error. These are referred to as census geocoding errors of inclusion. Results from the 1990 PES show that geocoding error is highly clustered. About 72 percent of the census geocoding errors were found in less than 3 percent of the PES sample block clusters. It seems that this is a clear example of a Deming principle, the so called "80-20" rule which states that in most cases "80 percent of the benefits are realized by solving 20 percent of the problems."

Targeting Methodology

The proposed plan for the 2000 A.C.E. is to target the extended search in the surrounding blocks of 20% of A.C.E. block clusters. Based on the 1990 PES experience we developed a well defined targeting criterion that when applied will result in the selection of A.C.E. block clusters with superior payoff. From the 1990 PES we learned that one reason for census omissions and erroneous enumerations was census geocoding error. This type of census omissions and erroneous enumerations will benefit the most from an extended search in surrounding blocks. *The criterion is to identify TES block clusters on the basis of independent listing unmatched housing units with a nonmatched census address. On the E-sample side the criterion is the number of housing units geocoded erroneously in the E-sample block.*

III. 2000 A.C.E. SAMPLING PLAN

An empirical simulation was designed and performed to assess the effect of alternative TES plans on the DSE and its variance. We used the 1990 PES data base for the simulations. The results are conditional on the 1990 PES experience. Although it is important to note that there are many differences between the 1990 PES and the 2000 Census A.C.E., the simulation results provide the basis for discriminating between alternative TES sampling plans. The TES sampling plans simulated fall into two categories; certainty selection and a combination of certainty and probability sampling. Certainty samples ranging in size from 5 to 20 percent were simulated. All these samples yielded more reliable DSE's compared to not doing any search but the DSE estimates differed from the 1990 DSE with 100 percent search. The difference results from "lack of balance." The reliability of the DSE based on a 20 percent TES certainty sample is very close to the precision of the 1990 DSE which was based on a full PES sample extended search. However, a small difference in the DSE estimates is still present. It is a very difficult task, perhaps impossible, to design a balanced certainty sample. To compensate for this, we developed several plans based on a combination of certainty and probability sampling. For these sampling plans, half of the TES sample was selected with certainty (Targeted) and the remainder was selected using a systematic sampling scheme. These sampling plans produced more consistent results at the expense of increased variance. Based on the simulation results, we developed the following sampling plan for implementation in 2000.

Certainty Sample

Five percent of clusters with the most census geocoding errors and independent listing address nonmatches . Five percent of clusters with the most weighted census geocoding errors and A.C.E. address nonmatches. In addition, all relisted clusters were included in sample.

Probability sample

A systematic sample from the remainder clusters with at least one census geocoding error or an independent listing unmatched address. The number of clusters in the sample will be determined so the total numbers of block clusters in sample is equal to 20 percent of the A.C.E. sample.

A.C.E sample clusters in List/Enumerate areas are out of scope for TES sample selection. List/Enumerate clusters will be handled through special procedures. Special procedures were developed as needed for clusters with high person nonmatch and census geocoding error rates.

SAMPLING OPERATIONS

Results from the initial housing unit matching operation were used to identify the TES sample. Housing unit matching consist of several operations. These are:

- Computer match - Addresses in the DMAF extract are computer matched to addresses in the A.C.E. Independent Listing.
- Clerical match - For this operation the search area is limited to the sample block cluster.
- Housing Unit Follow up - Results from the computer and clerical match operations are used to identify cases to go to the field for follow up. The goal of this operation is to create an accurate inventory of all housing units in the block cluster.
- After Follow up Coding - Using the information collected during field follow up housing units are assigned one of several codes.

For TES sample selection we are interested in three types of housing units, these are:

- CI - The A.C.E. housing unit existed as a housing unit and is correctly geocoded in the block cluster. An address corresponding to the housing unit is not found in the census.
- UI - Not enough information to determine the match status of the housing unit with certainty.
- GE - The census housing unit existed as a housing unit but is incorrectly geocoded in the block cluster. The housing unit is a geocoding error.

TES sample blocks were identified based on the number of housing units coded as GE, CI, and UI. The total number of housing units in these three categories were obtained for each block cluster. The sampling plan was implemented to select the TES sample. The search area is expanded for the block clusters selected in sample.

During the TES operations, housing units in the search area of a given TES cluster are searched for census omissions and erroneously enumerated housing units. If the census address corresponding to the unit is found in the surrounding blocks but was not included in the census then the persons in the housing unit are searched for duplication in the block where the housing unit is and if not found they are coded correct enumerations. If the household is determined to be a duplicate then the E-sample persons are coded erroneous enumerations.

The following information was available for each A.C.E. block cluster once the housing unit after followup clerical matching operation was completed:

a. number of P-sample whole household non-matches - correct A.C.E. housing units that did not match to a census address; and

b. number of E-sample geocoding error - housing units confirmed to exist outside the A.C.E. block cluster within the search area.

These two pieces of information can be combined or used individually to develop a set of reasonable criteria by which to select the TES block clusters. In 1990 the number of whole household nonmatches and E-sample geocoding error were highly clustered with the majority of the block clusters having none. We speculate that this will be the case in 2000 which will facilitate targeting. The criteria are the sums of unweighted and weighted a. and b.

The selection of TES block clusters is a combination of targeting and probability sampling. Relisted clusters and clusters where the matching operation was not performed in time for TES sample identification were included in the TES workload.

IV. Methodology

The definition of the search area concept has been the subject of research since the 1990 PES. Woltman (1994) examined the effect of a search area definition limited to the PES block cluster under certain conditions. This scenario can be described as zero or no extended search at all. Thus, the TES proposal should be viewed as a compromise between the 1990 PES search operation (or 100 percent search) and the 1995 Census Test PES zero or no extended search operation. His work was analytical in nature, using Taylor series to approximate the variance of the DSE. He concluded that for high match and correct enumeration rates the use of an extended search area results in dual-system estimates that are 10-15 percent more reliable than for a search area limited to the PES block cluster. Griffiths (1997) examined the effect

of a limited search area definition on the DSE and the coefficient of variation of the DSE using the 1990 PES data. He concluded the following:

- Averaged over the 357 poststrata, the increase in estimated standard error was 60.4 percent; the median increase in standard error over the 357 poststrata was 24.7 percent.
- The average increase of the dual-system estimate was 1.81 percent, the median increase was 1.50 percent. For the nation, the DSE was 1.94 percent greater.

All of the above results are conditional on the 1990 PES sample realization. The large difference between the analytical and empirical results is of particular interest. Griffith and Woltman were not able to reconcile or explain this difference. So far, research in this area has been limited to the size or the definition of the search area.

A. TES Sample Selection Plans

The TES operations determine the exact block location of the housing units that were identified as census geocoding error. This operation will be performed concurrently with A.C.E. interviewing. The selection criteria, described in Section II, are based on information that will be available from the housing unit matching operations. Housing unit matching will be performed prior to A.C.E. interviewing and will provide the geocoded errors (E-sample side) and address non-matches (P-sample side) that will be used to determine how many housing units might contain people subject to TES. For sample selection purposes, these are called "Interesting Housing Units" (IHU's) as they are the ones that will benefit from extended search. The IHU is the measure of size used for sample selection.

The sampling selection schemes are categorized as follows:

1. Certainty selection – This is a cut-off sample procedure. Block clusters were sorted by the sampling criterion in descending order. The top X% of blocks were selected for search with certainty. A negative aspect of this sample selection method is that it has the potential to increase the "balancing error", above what may already exist in the PES sample. Balancing error is the result of inconsistent treatment applied to P and E samples and introduces additional bias into the DSE. Therefore, the resulting DSE's are conditionally biased.

2. Certainty selection plus a systematic sample – This sampling selection method is a combination of certainty and probability sampling. The top X% of blocks were selected with certainty, and a systematic sample was selected among blocks not chosen with certainty. This method results in more consistent estimates at the cost of increased variance due to increased weight variation.

Results were tabulated and are available for all simulated sampling plans. However the analysis is concentrated on three of the plans. These are referred to as Plan B, D, and G. A description of the plan follows.

Plan B – the 5% of clusters with the most Interesting Housing Units, unweighted, were selected with certainty and the remaining 15% of the sample was selected systematically from all other clusters.

Plan D – the 5% of clusters with most Interesting Housing Units, based on both a weighted and an unweighted count, were included with certainty and the rest of the sample was selected systematically.

Plan G – 5% of clusters were selected with certainty based on weighted Interesting Housing Units, plus a systematic sample.

B. DSE Estimator

A simple expression for the DSE estimator is as follows:

$DSE = (C^*)\,(CE / E) / (M / P)$, where:

C^* = census data defined persons
CE = weighted estimate of correct enumerations
E = weighted E-sample total
M = weighted estimate of P-sample matches
P = weighted P-sample total

To implement the TES we needed a way to incorporate the people found in surrounding blocks into the DSE formula in a way that reflected the selection criteria and the sample design. Two estimators were developed for the component pieces C, E, P and M. C^* comes from the census and is not affected by the P- and E-samples or TES. Since the 2000 TES strategy calls for searching for particular addresses that are unmatched, we can define a person in such a housing unit as being a "TES Person". TES Persons are found in TES block clusters as well as those block clusters not selected for TES. Likewise, non-TES persons are in both types of block clusters.

Estimator 1

The TES sample results are only used to estimate the

number of matches and correct enumerations. The E and P sample total estimates are tallied using the A.C.E. weights without any TES adjustment. The weighted estimate of matches and correct enumerations does take into account the TES sampling. Assume block A is in a 20 percent TES sample, so the TES take-every or weight is 5. The PES weight was multiplied or adjusted by 5 to get the estimate of correct enumerations. The weighted estimates of E and P sample totals do not take into account the TES sampling. Thus, the E and P sample estimates of total population are fixed given the PES sample.

Estimator 2

The TES sampling weights are used to estimate all four sample components of the DSE. Therefore, TES persons in block clusters that were selected in the TES sample were weighted to produce the estimated components of the DSE. Thus, TES persons in blocks not selected in the TES sample are not tallied in the estimation of E and P sample totals. These non TES-sample people are accounted for by adjusting the PES weights by the TES sampling weights. This estimator takes advantage of the positive correlation between the estimate of correct enumerations (or P-sample matches) and the E sample (or P sample) estimate of total. As a result, estimator 2 is more reliable than estimator 1.

The CE and M components both have the same definition under each estimator. Anyone who was found in a surrounding block is weighted by their cluster's TES weight times the PES sample weight, in effect counting them that many times. For a cluster selected with certainty, the probability of selection, and hence the TES weight, is one. For a TES sampled cluster, the TES weight used is the reciprocal of the probability of selection.

Estimator 2 is not only superior from a results standpoint, but a more logical expression of the Dual System Estimation procedure. The "dual system" of the DSE is the calculation of the correct enumeration probability on the E-sample side and the match probability on the P-sample side. Each person in the E- or P- sample should have a probability of 0 or 1 in most cases of being either a correct enumeration or match. Under Estimator 1, people would count as multiple correct enumerations or matches, even though they represent just one E- or P- sample person in the sample total. Under Estimator 2, multiple correct enumerations or matches for one person correspond to the same number of replications in the P- and E-samples.

C. TES Simulations

This section describes in detail an empirical simulation performed to assess the effect of alternative TES plans on the DSE and its variance. We used the 1990 PES data base for all the simulations. Therefore, the results are conditional on the 1990 PES experience. Note that there are many differences between the 1990 PES and the 2000 A.C.E. designs. The main difference between the two procedures is the handling of movers during the estimation phase. The simulations **do not** address this issue. We did not change the match probabilities used for movers in 1990 to reflect the 2000 mover formula. The results of these simulations only reflect what would have occurred in 1990 had we used the TES procedures as described in section II. The results should be used as a barometer to compare the benefits associated with each alternative and not to predict a specific performance in 2000. The TES alternative plans were simulated by modifying the 1990 PES match and correct enumeration results to what would have happened under the 2000 A.C.E. TES plan. The following steps were implemented:

1. Recoding P-sample matches that were matched to persons in E-sample surrounding blocks as non-matches. Operationally, this means identifying records whose E-sample indicator is equal to "3" (surrounding block match) to a non-match and changing its match probability to zero.

2. Recoding E-sample persons in households that were a whole census household nonmatch found in a surrounding block and coded as correct enumerations to erroneous enumerations. Operationally, we would find geocodes (person found in a surrounding block) and change the associated probability of correct enumeration to 0.

3. Changing E-sample persons who are duplicated in E-sample surrounding blocks to having zero duplicates, which will usually change their probability of correct enumeration to 1, to reflect that without a surrounding block search such persons would not be found to be duplicates, but would be determined to be correct enumerations.

To simulate the TES plan, we change the match and correct enumeration probabilities back to what they had been in 1990 for the clusters that were selected for the TES sample. This process involves the following steps:

- Changing the match probability (P-sample) and probability of correct enumeration (E-sample) back to 1.

- Adjusting the final weights to account for TES sampling.

After simulating the 2000 TES methodology, we calculated 1990 DSE's and variances. Jackknife variance estimates were calculated using the SAS software.

V. Analysis and Results

Because the methodologies of the P- and E-samples are not exactly the same, it is possible, in fact likely, that the effect of doing an extended search will be different under the two samples. Simulating the effect of limiting the search area to the A.C.E. block cluster found that the direct DSE of total population was 1.5% higher than the 1990 DSE with full Surrounding Block Search (see Table 1, "Comparison of Certainty Only and Certainty with Sample TES Plans"). As we expected, simulations of certainty samples ranging in size from 5 percent to 20 percent show that as the TES sample size increases the difference between the estimated DSE's approaches zero. The larger the sample, the closer the DSE got to the 1990 DSE under full Surrounding Block Search and the smaller the variance. However, the DSE's show the effect of varying degree of balancing error.

It is a very difficult task, perhaps impossible, to design a balanced certainty sample. To compensate for this, we developed various plans that included a probability sampling component. Under these sampling plans, part of the TES sample was selected with certainty ("Targeted") and the remainder was selected using a systematic sampling scheme. A systematic sample was selected from the remaining block clusters not selected in the certainty sample. Plan "B", in which 5% of clusters were selected with certainty and 15% of additional clusters were selected from a 1-in-9 sample of the remaining clusters (and weighted), produced a DSE that appears consistent with the 1990 DSE. Simulation results of other sampling plans also showed that the resulting DSE's are very close to the 1990 DSE.

TES sample selection plans that include probability sampling produced more consistent results at the cost of increased variance. The average variance from five runs of Plan B was substantially less than that of doing no Surrounding Block Search, but was greater than the variance of even the 5% Certainty Plan. Our research indicated that much of this difference was due to the effect of clusters that had been assigned high weights in the cluster sample (i.e. the PES Weight.) We developed a "hybrid" selection criteria, a combination that included clusters based on their number of weighted IHU's along with the number of unweighted IHU's (Plan "D") or just the number of weighted IHU's without regard to the number of unweighted ones (Plan "G").

The average (from five sample realizations) variance under Plans D and G were only slightly higher than those under full Surrounding Block Search, and much smaller than under Plan B. Plans D and G also show much smaller maximum variances than Plan B. These two plans also produced more consistent DSE's. Large variance estimates for some groups under Plan B are reduced with Plans D and G.

Estimator

Both of the estimators used produced consistent DSE's compared to 1990 under all sampling plans involving a probability sampling component. The choice of one or the other should be driven by variance reduction. In general, estimator 2 is more reliable than estimator 1 for the largest population groups. Under Plans D and G for smaller population groups, the results are mixed.

Estimator 2 also, in our view, makes more sense from a logical standpoint. Correct Enumerations and Matches are supposed to be used as percentages of the total E- and P-sample persons. Theoretically, a group could show a match or correct enumeration rate of more than 100%.

References

Hogan, Howard (1993), "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association*, 88, 1047-1058.

Wolfgang, Glen (1999), "Request for Dress Rehearsal Surrounding Block Files for Accuracy and Coverage Evaluation Research", March 29, 1999, memorandum for Robert W. Marx, U.S. Bureau of the Census.

Woltman, Henry F., "Increase in Relvariance of DSE Without Surrounding Block Search", June 13, 1994, memorandum for Distribution List, U.S. Bureau of the Census.

Griffiths, Richard, "Search Area Definition for the Dual System Estimation", October 31, 1997, memorandum for Elizabeth A. Vacca, U.S. Bureau of the Census.

* This paper reports the results of research undertaken by Census bureau staff. It has undergone a 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.