# Targeted Extended Search in The Accuracy and Coverage Evaluation of the Census 2000

Glenn Wolfgang, Phawn Stallone, and Tamara S Adams,
Glenn Wolfgang, Bureau of the Census, Washington, DC, 20233

Key words: Surrounding block search, dual system estimation, geocoding error

## 1. Introduction

A goal of Census 2000 was to enumerate each person once at their residence on Census Day, April 1, 2000. The Accuracy and Coverage Evaluation (A.C.E.) estimated the number of persons missed or erroneously enumerated in the census. One step in the estimation process involved matching, in survey sample areas (clusters), persons enumerated in the A.C.E. survey to persons enumerated in the census.

The Targeted Extended Search (TES) extended the search beyond the cluster for missed or erroneous enumerations attributable to geocoding error. Geocoding error is the incorrect assignment of block and cluster identification code (geocode) to a housing unit. TES was a surrounding block search developed to enhance efficiency in field and processing workloads and to reduce the effects of geocoding error in estimation. Surrounding block search operations were used and developed in past census coverage evaluations. TES reduced dual system estimate variances inflated by census geocoding errors and provided robustness to A.C.E. geocoding errors.

This paper provides a brief description of TES procedures and some preliminary results on how TES effects on geocoding errors, reflected in matches and correct enumerations, relate to common variables.

## 2. Accuracy and Coverage Evaluation

The A.C.E. sample consisted of block clusters; large block clusters were subsampled. There were 11,303 A.C.E. clusters in the 50 states and District of Columbia. The addresses and people enumerated by the census in the A.C.E. clusters comprised the E sample, which was used to measure errors in the census data. The A.C.E. also independently listed the addresses and persons in the sample areas. Those persons comprised the P sample, which was used to measure people the census missed.

Data collection and processing began with the housing unit phase, in which addresses in the sample were listed and confirmed. Processing staff matched addresses independently listed by the A.C.E. to census housing unit addresses in the Decennial Master Address File. An address found in both sources was called a match; an address not found in both was a nonmatch. Staff conducted a housing unit field follow-up to confirm the existence of nonmatched housing units and to resolve other incomplete housing unit information.

In the person phase, A.C.E. field staff conducted independent interviews to obtain data on Census Day residents at the addresses listed in the housing unit phase of A.C.E. The people listed in the A.C.E. housing units comprised the P-sample. The people enumerated by the census in the A.C.E. clusters comprised the E-sample. Staff matched P-sample people to census enumerations, then followed up in the field and coded any cases needing additional information.

Other reports provided greater detail on A.C.E. and prior census coverage evaluations. Hogan (1993) reported on both analyses and procedures for the 1990 census. Hogan (2000) described application of theory in A.C.E. Childers (2001) described the A.C.E. design. Adams, Barrett, and Byrne (2001) summarized procedures for A.C.E. operations.

## 3. Targeted Extended Search

The TES involved searching outside the sample cluster for persons or housing units that were likely affected by geocoding error. The area in which A.C.E. looked for matches and correct enumerations was called the search area. The block cluster was the search area for non-TES cases. For TES cases, the search area was extended to the first ring of surrounding blocks, which included any block touching the A.C.E. cluster at any point, even if only at a corner point. The extended search was targeted to selected clusters and selected households in those clusters which showed potential for geocoding error, as measured by A.C.E. housing unit phase results.

The TES differed from the 1990 surrounding block search primarily in its focus on geocoding error. The 1990 procedures extended the search area for matches,

---

The authors are mathematical statisticians in the Decennial Statistical Studies Division of the U.S. Census Bureau. 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.

correct enumerations, or duplicates for all clusters. In contrast, the TES targeted clusters and specific households where the effort would be most beneficial for handling geocoding error. Another difference was that the search area was limited to one ring of surrounding blocks for the TES. In 1990, a second ring of surrounding blocks (comprised of blocks not touching the cluster but not more than one block away) was searched for update/leave type of census enumeration areas, and the entire address register area was searched for list/enumerate clusters. The TES was designed to improve the operational efficiency and the quality of the extended search over 1990 procedures.

Extended search procedures involved both clerical matching of data records and field follow-up visits. The TES field follow-up was conducted about the same time as A.C.E. person interviewing. In clusters selected for the TES, the field staff canvassed the cluster or surrounding blocks to locate census housing units identified as potential geocoding errors. If the housing unit was found in the search area, the data of persons enumerated at that unit were reviewed to determine if those persons were correctly enumerated or duplicated in the census.

Two types of census geocoding errors and one type of A.C.E. geocoding errors occurred:

- **Census errors of inclusion** occurred when a housing unit physically located outside a sample cluster erroneously had an in-cluster block number on census records. In this situation, a field search for the address limited within the cluster would have failed to confirm that it was correctly enumerated by the census. As a result of the TES field follow-up finding the address in the search area, the people enumerated at that address, who would have been census errors of inclusion and coded as erroneous enumerations were recoded as duplicates or correct enumerations.
- **Census errors of exclusion** occurred when a housing unit physically located within a sample cluster erroneously had a not-in-cluster block number on the census record. In this situation, persons listed in the A.C.E. may not have been matched to census persons enumerated at that address. When the search was extended to census records in the surrounding blocks and clerks found matching persons, P-sample persons who would have been census errors of exclusion and coded as nonmatches were instead recoded as matches.
- **A.C.E. geocoding errors** occurred when a housing unit physically located outside a sample cluster had an A.C.E. address record with an in-sample block cluster number. In this situation, as with census errors of exclusion, persons listed in the A.C.E. may not have been matched to census persons enumerated at that address. When the search was extended to census records in the surrounding blocks and clerks found matching persons, P-sample persons who would have been A.C.E. geocoding errors and coded as nonmatches were instead recoded as matches.

The potential for geocoding error was determined in the housing unit matching and follow-up preceding person interviewing. Nonmatched E-sample addresses confirmed by field follow-up to exist as a housing unit outside the sample cluster point to census errors of inclusion. Nonmatched P-sample addresses suggest census errors of exclusion or A.C.E. geocoding errors. A cluster's sum of these nonmatched addresses in both E and P samples was the measure of the potential geocoding error used for targeting clusters for TES.

The targeting of potential geocoding error cases was designed at both cluster and household levels:

- **Cluster Targeting and Sample Selection** – Targeting clusters reduced the number of clusters selected for extended search by nearly 80 percent. First, TES selection included with certainty 62 clusters for which housing unit matching was too delayed to ascertain the geocoding status of the census units in those clusters. List/enumerate clusters, for which census data was not available in time for TES field follow-up, were excluded from TES. From the remaining clusters, 1,088 with the highest weighted and unweighted measures of geocoding error potential were selected for TES with certainty. Another 1,089 clusters were selected by sampling from those with a non-zero count of potential geocoding error.
- **Address Targeting** – Rather than processing every address in TES clusters, the operation targeted households with characteristics of geocoding error. Specifically, TES targeted addresses as follows:
  - P sample – During person matching, matching staff searched the census data of surrounding blocks only for whole-household nonmatches (that is, residents of households where all were nonmatched) in any P-sample nonmatched housing unit. Matches resulting from these searches were attributable to census errors of exclusion or A.C.E. geocoding errors. We also limited our surrounding block search in urban areas to the block in which a matching census address was found. In clusters with one or more non-city style address, we searched in all surrounding blocks.
  - E sample – During the time of person interviewing, a TES field follow-up confirmed whether E-sample addresses coded as potential census geocoding error were indeed located in the search area. Only whole-household nonmatches

at addresses confirmed to be in the search area were attributed to census errors of inclusion and recoded as correct enumerations. A duplicate search for any people coded outside the cluster was limited to the block in which the housing unit was located in TES field follow-up.

## 4. Effect of TES on Estimates

The TES identified matches and correct enumerations within the search area in housing units with geocoding errors. Without TES, the number of both would have been be lower. A few duplicates were also identified but remained erroneous enumerations and did not affect the number of correct enumerations. The dual system estimate (DSE) formula shows the role of matches (M) and correct enumerations (CE), expressed as the proportions (M/P and CE/E) of their respective samples, along with the number of data-defined census persons excluding late census adds and whole-person imputations (DD):

$$DSE = DD * (CE/E) / (M/P)).$$

One can see from the DSE formula that if matches and correct enumerations both changed about the same amount with the increased search area, while E and P sample totals remained about the same, the expected value of the DSE will not be affected. Because A.C.E. sampling was random, any given case of census geocoding error affecting the sample should be as likely erroneously included as erroneously excluded. As long as the search areas for P-sample and E-sample cases are kept the same, census errors of inclusion and census errors of exclusion should be equal and balance each other in the sense of changing matches and correct enumerations at the same rate. Mulry and Spencer (1991) discussed balance.

The A.C.E. geocoding errors, on the other hand, were not balanced with anything else, but were expected to be few. Like extended searches in previous censuses, the TES was designed to provide robustness to A.C.E. geocoding error (Navarro and Olson, 2001). Navarro and Olson noted that A.C.E. geocoding error might explain any differences observed between rates of matches and correct enumerations in surrounding blocks. Using data from additional field operations, Adams and Liu (2001) evaluated the potential lack of balance in the TES and concluded that the TES provided robustness against A.C.E. geocoding error.

Navarro and Olson also reported preliminary analysis of the TES impact on variances, "At the post-stratum level the average weighted improvement is 33 percent. . . . . So there can be little question that TES makes DSE estimates more precise, . . . ."

## 5. Limitations

There were certain limitations in the results presented in this paper. Several were computational shortcuts with negligible impact on test results and interpretations that permitted the efficiency and versatility needed to conduct a wide range of analyses.

- Inmover data were not used as in official dual system estimates. See Haines (2001) for a description of the conditions and methods for using inmover data in official estimates. See Davis (2001) for the official statistics. The number of matches was slightly inflated in all groups compared – with little effect on group differences which were tested.
- Match statistics excluding surrounding block matches were not precisely the same as match results if no TES were done, because follow-up and estimation might affect the residence status or weighting of the persons involved.
- Standard error computations in these analyses were simplified and did not take all levels of the sampling into account. We expected only a trivial impact on variances due to computing methods; we expected no impact on test results or conclusions.

## 6. Analysis Methods

In addition to reporting some general results, this study analyzed differences among percentages of surrounding block matches and percentages of surrounding block correct enumerations. Those statistics were computed within subgroups of the sample defined by levels of post-strata or other operational variables. Specifically, surrounding block matches were matches assigned only due to TES operations and the denominator for the percent of surrounding block matches was the number of P-sample persons in the subgroup. Surrounding block correct enumerations similarly were assigned only in TES operations. The denominator for the percent of surrounding block correct enumerations was the number of E-sample persons in the subgroup. Full sample and TES weighting of these numbers was used. Groups with high percentages of surrounding block matches or surrounding block correct enumerations would provide insight on conditions associated with geocoding error.

Stratified Jackknife methods were used to compute variance estimates for the statistics. Significance testing used a Bonferroni criterion, controlling the joint error probability, set equal to 0.10, for all tests conducted among the levels of one variable. In addition, tests with levels based on less than 100 person records were avoided, either through collapsing with other levels or by dropping that level from tests.

## 7. Results

A few overall results showed the scope and impact of the TES. Table 1 shows that due to targeting, which focused efforts on potential geocoding error cases or a representative sample of those cases, only 6 percent of weighted P-sample persons and 3.2 percent of weighted E-sample persons were involved in TES processing.

The size of the TES in the P sample was much larger than in the E sample largely because many P-sample nonmatches selected for TES were not geocoding error cases. In contrast, almost all erroneous enumerations selected for TES resulted from geocoding error.

If 1990 extended search procedures had been used, the size of the operation would have been about double that of the TES. For the P-sample, all persons not matched in the sample clusters, about 12 percent (See Table 2 Matches In Sample Clusters) would have needed extended search. For the E-sample, the 7.6 percent of persons not confirmed as correct enumerations in the sample clusters (See Table 3 below) would have needed extended search.

**Table 1 Size of the TES in P and E Samples**

| | P Sample Weighted | E Sample Weighted |
|---|---|---|
| Total Sample | 258,547,382 | 264,578,862 |
| non TES | 243,077,600 | 256,034,032 |
| TES | 15,469,782 | 8,544,830 |
| TES(%) | 6.0% | 3.2% |

Table 2 shows that the surrounding block matches made up about 3.9 percent of the P sample, using nonmover and outmover data without any inmover data information. Navarro and Olson (2001), using inmover data as in official computations, report a 3.8 percent. These TES cases matched in surrounding blocks represent the sum of geocoding error due to census errors of exclusion and A.C.E. geocoding errors. The two types of error cannot be distinguished using information available to this study.

Table 3 shows that surrounding block correct enumerations made up about 2.9 percent of the E sample. These TES cases coded correctly enumerated in surrounding blocks represented geocoding error due to census errors of inclusion.

The difference in surrounding block matches and surrounding block correct enumerations was explained as due largely to A.C.E. geocoding errors (Adams and Liu, 2001)

**Table 2 Matches in Surrounding Blocks**

| | Weighted Number | Weighted Percent |
|---|---|---|
| Total P Sample | 258,547,382 | 100.0 |
| Matches: | | |
| Total | 237,401,214 | 91.8 |
| In Sample Clusters | 227,399,141 | 87.9 |
| In Surrounding Blocks | 10,002,073 | 3.9 |

**Table 3 Correct Enumerations in Surrounding Blocks**

| | Weighted Number | Weighted Percent |
|---|---|---|
| Total E Sample | 264,578,862 | 100.0 |
| Correct Enumerations: | | |
| Total | 252,096,238 | 95.3 |
| In Sample Clusters | 244,387,951 | 92.4 |
| In Surrounding Blocks | 7,708,287 | 2.9 |

As Table 4 shows, very few TES duplicates were found.

**Table 4 Duplicates in Surrounding Blocks**

| | Weighted Number | Weighted Percent |
|---|---|---|
| Total E Sample | 264,578,862 | 100.00 |
| Duplicates: | | |
| Total | 1,852,499 | 0.70 |
| In Sample Clusters | 1,759,313 | 0.66 |
| In Surrounding Blocks | 93,186 | 0.04 |

Results from comparing statistics are presented below in tables displaying variable level names with level numbers assigned for reference in the third column, values for the statistic of interest (in columns headed “percent”), the stratified jackknife standard error (in columns headed “s.e.”), the weighted percent of persons contributing data to the analysis (in columns headed “n”), and a list of the level numbers with which a significant difference was found (in columns headed “differs from"). The criterion t value that applies in that table was noted

below each table. The levels were arranged in ascending order of percent to help display data patterns.

Post-stratification variables were important to A.C.E. estimation. Tables 5-8 show results for analyses of two post-stratification variables.

One result was evident in analyses of Tenure. Home owners had a lower percent not matched in the whole P sample (6.1 as compared to 13.1 for non-owners; see Wolfgang, Davis, and Stallone, 2001). Likewise, they had a lower percent of matches in surrounding blocks. In other words, more matches in surrounding blocks were found for the group where there were more nonmatches in which they might be found.

**Table 5 Percent of Matches in Surrounding Blocks by Home Ownership**

| Tenure | Percent | Differs from | s.e. | n |
|---|---|---|---|---|
| 1: Owner | 3.3 | 2 | 0.2 | 69.8 |
| 2: Non-owner | 5.1 | 1 | 0.6 | 30.2 |

Note: Criterion for levels to differ was | t | > 1.645

Similarly, E-sample home owners had a lower percent of erroneous enumerations (Feldpausch, 2001) and a lower percent of correct enumerations in surrounding blocks. More correct enumerations in surrounding blocks were found for the group where there were more erroneous enumerations in which they might have been found.

**Table 6 Percent of Correct Enumerations in Surrounding Blocks by Home Ownership**

| Tenure | Percent | Differs from | s.e. | n |
|---|---|---|---|---|
| 1: Owner | 2.7 | 2 | 0.2 | 69.7 |
| 2: Non-owner | 3.4 | 1 | 0.3 | 30.3 |

Note: Criterion for levels to differ was | t | > 1.645

However that relationship observed for Tenure between overall statistics and surrounding block statistics did not show up consistently in other variables. In fact, only MSA/TEA, among other post-strata variables, yielded more significantly different test results than were expected by chance.

Metropolitan Statistical Area (MSA) designates cities for statistical purposes. Type of Enumeration Area (TEA) designates the method of data collection adopted for an area. MSA size and Mailout/Mailback TEA (versus all others) were combined into one variable used for post-stratification.

For both surrounding block statistics, all levels of Mailout/Mailback type of enumeration area stood out with higher percentages of correctly enumerated in surrounding blocks than the level combining all other types of enumeration area. The correspondence between the TES and the general statistics did not show up as it did with Tenure.

**Table 7 Percent of Matches in Surrounding Blocks by Size of Metropolitan Statistical Area and Type of Enumeration Area**

| MSA/TEA | Percent | Differs from | s.e. | n |
|---|---|---|---|---|
| 1: All Other TEAs | 1.9 | all | 0.3 | 18.1 |
| 2: Small MSA & Non-MSA Mailout/Mailback | 4.0 | 1 | 0.4 | 20.2 |
| 3: Large MSA, Mailout/Mailback | 4.4 | 1 | 0.5 | 30.4 |
| 4: Medium MSA, Mailout/Mailback | 4.4 | 1 | 0.5 | 31.3 |

Note: Criterion for levels to differ was | t | >2.386

**Table 8 Percent of Correct Enumerations in Surrounding Blocks by Size of Metropolitan Statistical Area and Type of Enumeration Area**

| MSA/TEA | Percent | Differs from | s.e. | n |
|---|---|---|---|---|
| 1: All Other TEAs | 0.7 | all | 0.1 | 17.9 |
| 2: Large MSA, Mailout/Mailback | 2.8 | 1 | 0.3 | 30.2 |
| 3: Small MSA & Non-MSA Mailout/Mailback | 3.5 | 1 | 0.3 | 20.4 |
| 4: Medium MSA, Mailout/Mailback | 3.9 | 1 | 0.4 | 31.5 |

Note: Criterion for levels to differ was | t | >2.386

Generally, fewer than 10 percent of the tests in each of the eight other post-stratification variable tables were significant. Only three of the tables, all surrounding block correct enumeration analyses, had any significant differences, and in each of them there was but one group that differed from some but not all other groups: Females

aged 18-29 (3.5), Native Hawaiian or Pacific Islander (1.1), and the Midwest region (2.2).

That number of statistically significant differences was within the number expected by chance with the criterion level applied. Even though it was tempting to draw other conclusions about those few differences, it was possible that those results were primarily Type I error or simply not worth much attention.

Other operational variables were tested. A low percent of matches in surrounding block and a low overall percent not matched, as for home owners, were found for persons in subsampled clusters, single family homes, or proxy-response households. Surrounding block statistics related inconsistently or not at all to imputation status of post-stratification variables, mover status, or household size.

## 8. Conclusions

A few conclusions may be drawn from this study.

- Targeting kept the operation small.
- A.C.E. geocoding error may explain the overall difference between numbers of surrounding block matches and surrounding block correct enumerations.
- Based on Navarro and Olson (2001), TES substantially reduced dual system estimate variances.
- There were few TES duplicates.
- Sometimes, but not consistently, a large overall percent not matched appeared to coincide with a large percent of matches in surrounding blocks, or percent erroneous enumerations with percent of correct enumerations in surrounding blocks.
- There were few indications of a relationship between post-stratification or other operational variables and the surrounding block statistics. None of the variables tested provided a means to better understand or manage geocoding error.

## 9. References


Adams, T., Barrett, D., and Byrne, R. (2001). "Operational Plan for Accuracy and Coverage Evaluation (A.C.E.) for Census 2000," DSSD Census 2000 Procedures and Operations Memorandum Series S-TL-06, U.S. Census Bureau, Washington, D.C.

Adams, T. and Liu, X. (2001). "Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report Number 2: Evaluation of Lack of Balance and Geographic Errors Affecting Person Estimates" DSSD Census 2000 Procedures and Operations Memorandum Series T-13, U.S. Census Bureau, Washington, D.C.

Childers, D. (2001). "The Design of the Census 2000 Accuracy and Coverage Evaluation (A.C.E.)" DSSD Census 2000 Procedures and Operations Memorandum Series S-DT-01, U.S. Census Bureau, Washington, D.C.

Davis, P. (2001). "Accuracy and Coverage Evaluation: Dual System Estimation Results," DSSD Census 2000 Procedures and Operations Memorandum Series B-9*, U.S. Census Bureau, Washington, D.C.

Feldpausch, R. (2001). "Census 2000 E-Sample Erroneous Enumerations," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.

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

Hogan, H. (2000). "Accuracy and Coverage Evaluation: Theory and Application," Internal document, U.S. Census Bureau, Washington, D.C.

Haines, D. (2001). "Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Dual System Estimation (U.S.) -Re-issue of Q-37," DSSD Census 2000 Procedures and Operations Memorandum Series Q-48, U.S. Census Bureau, Washington, D.C.

Mulry, M. and Spencer, B. (1991). "Total Error in PES Estimates of Population," Journal of the American Statistical Association, 86, 839-854.

Navarro, A. and Olson, D. (2001). "Accuracy and Coverage Evaluation: Effect of Targeted Extended Search," DSSD Census 2000 Procedures and Operations Memorandum Series B-18*, U.S. Census Bureau, Washington, D.C.

Wolfgang, G., Davis, P., and Stallone, P. (2001). "Accuracy and Coverage Evaluation Persons Not Matched in Census 2000," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.