THE 1987 POST-ENUMERATION SURVEY

Irwin Anolik, Bureau of the Census*
Washington D.C. 20233

ABSTRACT

This paper discusses the methodology and results of the 1987 Post-Enumeration Survey (PES) in North Central North Dakota. The 1987 PES was a test in a sparsely populated rural area. Addresses in such an area are often unusable for matching. The main objective was to refine rural matching techniques in preparation for the 1990 Decennial Census evaluation. Production of census coverage estimates relies on an extensive system of matching between persons sampled in a PES and persons enumerated by the census. First, a computer typically matches most of the cases. Next, clerks review those cases not matched by the computer. This paper illustrates that a combination of computer and clerical matching can be effective in rural areas.

1. INTRODUCTION

The 1987 PES was a matching study conducted after the 1987 Census of North Central North Dakota. It was designed to gain experience in computer and clerical matching with rural addresses. Another objective was to evaluate the effectiveness and accuracy of the computer match using different versions of the matcher.

In the North Dakota test site, many addresses consist of a rural route and box number with no house number or street name. Blocks are sometimes irregularly shaped with "invisible" boundaries (e.g., an intermittent stream or a county line). An enumerator may list the wrong block or mistakenly include parts of neighboring blocks. This could lead to uncounted persons in missed housing units as well as duplicated persons in housing units counted more than once.

The 1987 PES involved a two-way match between persons sampled in the Rural PES (P sample) and persons enumerated by the census in PES sample blocks (E sample). Data from the P sample are used to estimate the number of persons missed in the census who should have been counted (gross undercount). Data from the E sample are used to estimate the number of persons incorrectly counted in the census such as duplicate enumerations and fictitious persons (gross overcount). The 1987 PES was primarily a matching study and did not include imputation and estimation activities.

2. SAMPLE DESIGN

The 1987 PES sample was drawn from the ten counties of the North Central North Dakota test site. The population in these counties was stratified using demographic data from the 1980 Census. Four strata were formed as follows:

Stratum 1: Prelist[1] blocks containing 0,1, or 2 housing units.

Stratum 2: List/Enumerate[2] (L/E) Address Register Areas (ARAs) with an American Indian population of ten percent or less.

Stratum 3: L/E ARAs with an American Indian population of more than ten percent.

Stratum 4: Prelist blocks containing 3 or more housing units. In this stratum, some blocks were combined to form block clusters with at least 6 housing units.

A sample of 36 ARAs from the L/E area and 62 blocks from the prelist area was selected. The targeted sample size was 1500 housing units. In the prelist areas, blocks were used as primary sampling units (PSUs). Because of the low population density in the L/E areas, we used ARAs instead of blocks as PSUs.

The L/E ARAs with large numbers of housing units were subsampled in an attempt to meet the targeted sample size and reduce interviewing workloads and costs. The subsampling defined compact cluster areas composed of blocks that would minimize travel for interviewers. In effect, blocks were subsampled from the originally sampled ARAs. This ensured overlap of the P sample and the E sample, and helped to determine if persons were counted in the census but missed in the PES.

3. FIELD ACTIVITIES

The field activities were address listing and interviewing, including the quality control checks on these activities.

3.1 Address Listing

The first phase of field activities for the 1987 PES was address listing. This produced an independent listing of addresses in all sample blocks. The listing phase of a PES is very important, particularly in a rural area. Addresses in such an area regularly consist of a rural route and box number with no house number or street name, and blocks are often bounded by unnamed roads. Thus the quality control (QC) check of the address listing takes on added importance.

As a quality control check in previous PESs conducted in urban areas, an administrative list of addresses was geocoded to specific blocks and compared to the address listings. For the 1987 PES an administrative list of addresses that could be compared to the address listings was not available. (Addresses are not geocodable to specific blocks in much of the test site). Therefore, the QC operation involved advance listing a sample of housing units in the PES sample. The advance listing was done by crew leaders and experienced interviewers prior to the regular address listing. After the block was listed by the regular interviewer, the QC clerk determined if the right block was listed and if all addresses were reported correctly.

An address listing book (ALB) failed QC if there were any discrepancies between the PES interviewer listing and the advance listing. Any ALB which failed QC was sent back to the field for rectification. Table 1 shows the breakdown of the quality control operation.

Table 1: Address Listing Quality Control Results

	No. of ALBs	Percent of ALBs
Pass	31	49.2
Fail	32	50.8
Total	63	100.0

During the QC operation, corrections were made to 13 (41%) of the 32 ALBs which failed QC. One block had to be relisted when the wrong block was originally listed. This indicates the kind of geocoding error that can occur in the census (see Section 1).

3.2 Interviewing

After completion of the Address Listing the next major field activity was interviewing. The 1987 PES interview obtained demographic data on all current residents, where they lived on Census Day, any alternate addresses (such as a college address), mailing address, and other related information on persons who lived at the address on Census Day.

The final outcome of the interviewing for all PES questionnaires checked into the Collection Office is given in Table 2.

During the first three weeks of interviewing, only interviews with household members were accepted. During the fourth week of interviewing, proxy interviews with nonhousehold respondents, such as neighbors or landlords, were permitted. The final few days of interviewing allowed for last resort data with whatever information the interviewer could obtain on the household. Fortunately, the need to collect last resort data never presented itself for the 1987 PES as we see in Table 2. Table 2 also shows a very low noninterview rate for the 1987 PES. Assuming high quality data, such a low noninterview rate is a desirable outcome and aids in controlling the error component associated with missing data - one of the eight main components of error generic to coverage measurements produced by post-enumeration surveys as pointed out by Hogan and Wolter (1988).

Table 2: Final Outcome of Interview

	No.	Percent	Percent of Occupied Housing Units
Complete Interview	1391	74.5	96.2
Vacant	420	22.5	NA
NI*-Refused	1	0.1	0.1
NI-Not at Home	0	0.0	0.0
NI-Other	1	0.1	0.1
Proxy	53	2.8	3.7
Last Resort	0	0.0	0.0
Total	1866	100.0	100.0

* Noninterview

A quality control check of the interviewing involved either telephone calls or personal visits to a sample of households to determine if the right household was interviewed and whether all the correct household members were included on the PES roster of names. Of the 253 work units[3], all passed QC. Obviously there was no evidence from the QC of any fabrication in the PES. Fabrication is another main component of error that the 1987 PES was apparently able to control successfully.

4. MATCHING

1987 PES matching was affected by two design decisions. One decision was to use a "PES B" procedure to determine match/nonmatch status. In this procedure, the PES interviewer lists all the persons living or staying in the housing unit at the time of the PES. The PES information for nonmovers is matched with the census. In-movers (persons who moved into the sample block between Census Day and the PES interview) are asked where they lived on Census Day. Their Census Day address is searched in attempting to match in-movers to the census. If their Census Day address is outside the test site, then the person is coded as being out-of-scope.

The major alternative to the PES B approach is called "PES A." The PES A procedure reconstructs the households as they existed at the time of the census. It attempts to obtain names and basic characteristics of persons who moved out (out-movers) between Census Day and the time of the PES interview. In either case the PES information is then matched with the census data. The difference between PES A and PES B involves people who move between Census Day and the time of the PES interview.

The PES B procedure was chosen for the 1987 PES because it reduces the need to get information from neighbors or from other non-household members as to who was living in the housing unit at the time of the census. However, it requires that in-movers give complete and accurate information on where they were living at the time of the census. This information is used in searching for the persons in the census listings at these former locations.

The second design decision affecting matching involved determining the extent of search. We decided to use an approach referred to as "correct address matching" which searches the census files where the person should have been enumerated in the census. The P-sample person is coded as a match when (s)he is enumerated at the correct census day address, as determined by census residency rules. Otherwise, a nonmatch is assigned. Depending on the type of enumeration area a search area was defined around each address. In the prelist areas, this search area was the block containing the address and two rings of blocks surrounding that block. In L/E areas, the search area was the block containing the address and the remainder of the ARA. Note that for movers, matching searches have to be made in non-sample areas.

4.1 Computer Matching

1987 PES interview questionnaires were keyed and the data were sent to headquarters where the PES files were prepared for computer matching. Similarly, census files were created for the same purpose. The census files included names, addresses, census processing data and demographic information.

The number of records in the PES and Census files is too large to consider all possible record pairs. The files are therefore partitioned into "logical blocks" so that comparisons are restricted to record pairs within each logical block. This blocking is implemented by sorting the two files on one or more variables. Such blocking variables ideally should have a large number of uniformly distributed value states and a low probability of reporting error. Blocking is a tradeoff between computation cost (examining too many record pairs) and false nonmatch rates (classifying record pairs as nonmatches because the records are not members of the same logical block).

The computer matching was done in a single pass in which the matcher initially "blocked" (i.e., sorted) on the following 2 variables: (1) Block Numbering Area (BNA), and (2) SOUNDEX of last name. The SOUNDEX procedure enables a variable such as surname to be phonetically encoded and allows matching despite minor spelling differences.

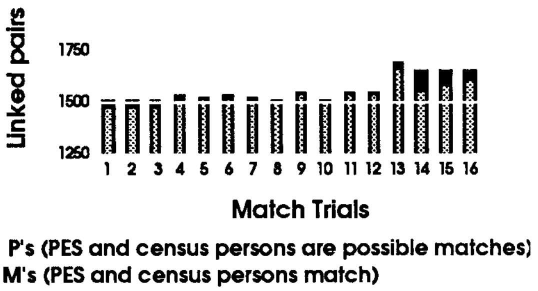
Other variables used for the computer match include the following:

1. Given name
2. Middle initial
3. Relation to head of household
4. Sex
5. Race
6. Marital status
7. Hispanic origin
8. Address
9. Phone number

The theory of computer matching, is discussed by Fellegi and Sunter (1969). Jaro (1986) and Winkler (1988) describe the application of the so-called Fellegi - Sunter algorithm to PES computer matching.

As part of the 1987 PES study, multiple trials of the computer match were performed using different combinations of variables and blocking schemes and different types of comparisons made for certain types of variables . {See Belin (1988) for more detailed discussion of such trials}. As figure 1 below illustrates, no appreciable improvements in the match rate occur until trial 13. This trial included, for the first time, using the first letter of the last name as a blocking variable, as opposed to the 4 character SOUNDEX of the last name. This blocking scheme was less restrictive and therefore yielded more matches, but more importantly, it did not sacrifice matching accuracy (see Section 4.2.3).

Figure 1. 1987 PES Match Results



A few additional trials with further refinements to the matcher did not yield much, as figure 1 illustrates. After a cursory examination of the output, trial 14 was chosen as the production match that clerks later reviewed (see Section 4.2).

The results of computer matching were encouraging, due largely to the refined blocking scheme discussed above and to the high quality of data on both the PES and Census files.

The overall computer match rate was 77.9%. This compares favorably with the 74.2% computer match rate obtained during the 1986 Test of Adjustment Related Operations conducted in Central Los Angeles County (Diffendal, 1988) and the 68.0% rate obtained for the 1986 Rural PES conducted in East Central Mississippi (Anolik, 1988). If we look only at nonmovers, the computer match rate for the 1987 PES jumps to 84.0%.

4.1.1 Extended Search Results

Census questionnaire information, including names, was keyed for the entire 1987 test site. Therefore, it was possible to detect geographic coding errors by computer as well as clerically. Such an automated extended search was, in effect, incorporated into the computer matcher by the use of BNA as a blocking variable (as discussed earlier). This enabled P-sample persons to be automatically matched to persons enumerated in different blocks within the same BNA. A total of 179 nonmovers were matched outside the PES sample block. Of these, 102 (57.0%) were matched by computer and 77 (43.0%) were matched clerically. Table 3 shows nonmover matches broken down by whether the match occurred within the block or outside the block. These results suggest the importance of defining an appropriate search area in a rural area since a number of matches can be found in surrounding blocks.

Table 3. Extended Search Results

	No.	Percent of Nonmover Matches	Percent of Nonmovers Total In-Scope
Matched Within Block	3205	94.7	90.5
Matched Outside Block	179	5.3	5.1
Nonmover Matches	3384	100.0	95.5
Total Nonmovers	3542	NA	100.0

Geographic errors in a rural site may be attributed to rural addresses and geography as discussed earlier. Postal delivery in such an area may also play a role. For instance people have a mailbox across the street from where they live (i.e., in another block). The geography of the mailbox may often be recorded on the address control file of the census instead of the location of the housing unit.

4.2 Clerical Review

The Clerical Review for the 1987 PES was completed by a clerical staff in Jeffersonville, Indiana. This was followed by a review by a more experienced staff, called the Special Matching Group (SMG), that ensures consistent and accurate matching results. All computer match forms were reviewed. Many of the nonmatches and possible matches were easily and quickly converted to matches by reviewing the persons in the household together. For instance, children from a previous marriage not matched because of inconsistent reporting of surnames can be matched when the parents are matched.

4.2.1 Review of Possible Matches

All possible matches were reviewed clerically and many were matched by examining PES and Census questionnaires.

Table 4 shows the total number of matched persons on the PES file broken down by computer matches and computer possible matches that were later clerically matched.

Table 4: Match Results for Combined Automated and Clerical Operation

	Number	Percent of Final Matches
Computer Matched Remained Matched	3086	84.1
Computer Possible Match-Clerically Matched	306	8.3
Total Matched[a] on PES File	3671	100.0

[a] The PES file includes nonmovers, PES B in-movers and PES A out-movers.

As we see from table 4, 84.1% of the persons ultimately matched on the PES file were initially matched by computer. An additional 8.3% were computer possible matches that were matched clerically. Computer matching thus linked together 92.4% of the cases that were ultimately matched.

4.2.2 Review of Nonmatches

All PES persons on the PES file not matched by computer were reviewed by the clerical staff and the Special Matching Group. The results of this clerical review are shown in Table 5.

Table 5: Results of Clerical Review

	Number	Percent
Total Computer Nonmatched	900	100.0
Matched Clerically	403	44.8
Matched by the SMG	182	20.2
Remaining Nonmatched	315	35.0

This table shows that 65% of the cases that were nonmatched to the census by the computer were matched during the clerical review.

4.2.3 Review of Computer Matches

All matches assigned by the computer were reviewed. Of 3,087 computer matches, only 1 was found to be matched erroneously. This error rate rate is extremely low. Refinements to the computer matcher and a more limited search area should play a role in keeping this error rate low in the 1990 PES.

4.3 Results of Matching at Alternate Addresses

To assist in matching to the census, any addresses at which a person may have been counted were recorded during the PES interview. Examples of such addresses include colleges, military bases, and second homes. For persons not living at the sample address on Census Day, their Census address was recorded. Results of matching at these alternate addresses will now be examined.

4.3.1 Results of Matching PES Movers

Persons reporting to have moved into a PES sample address between Census Day and the PES are called PES B in-movers. PES B in-movers are more difficult to match to the census because reported Census Day addresses can be incomplete or difficult to geocode to the census. Studies of other censuses have confirmed that persons moving at a time close to Census Day are at greater risk of being omitted from the census or of being enumerated at a subsequent address rather than at their correct Census Day address (Fay et.al., 1988). Table 6 shows the results of matching PES B in-movers.

Table 6. Results of Matching PES B In-movers

	No.	Percent of In-Scope Movers	Percent of Total
Matched at Reported Census Day Address	96	73.8	27.2
Out-of-Scope	223	NA	63.2
Nonmatched	26	20.0	7.4
Unresolved	8	6.2	2.3
Total Movers	353	100.0	100.0

As one might expect, many of these movers were out-of-scope or outside the test site at their Census Day address. Hence, they should not have been counted in the census. Table 6 shows that 73.8% of those cases reported as "in scope" were matched at their Census Day address. For these movers, both their sample address and their Census Day address were within the test site. Table 6 does not show that 66 PES movers were matched by computer to their PES sample address. These cases were either enumerated incorrectly at their PES sample address rather than at their correct Census Day address or they incorrectly reported their Census Day address in the PES.

4.4 P-Sample Matching

The results of matching the P sample are shown in table 7 for nonmovers and PES B in-movers.

Table 7. Summary of P-Sample Matching

	Number	Percent of In-Scope P Sample	Percent of Total
Matched	3480	94.8	89.3
Nonmatched	169	4.6	4.3
Out-of-Scope	223	NA	5.7
Unresolved	23	0.6	0.6
Total	3895	100.0	100.0

Most of the out-of-scope cases are persons with Census Day addresses outside the test site. In 1990, there will be a search area for these out-of-scope persons in the census, except for those who lived outside the country on Census Day.

We see from table 7 that 99.4% of the P-sample cases had their match/nonmatch status resolved without any field follow-up. Although we did not produce coverage estimates for this study, the high match rate (94.8% of in-scope P-sample cases) hints at a relatively low undercount for the 1987 test census. Since the test site provides a relatively easy area to enumerate, the high match rate supports what we would expect in such an area.

5. CONCLUSIONS

The 1987 PES was primarily a matching study in a sparsely populated rural area. Addresses in such an area are often unusable for matching. The main objective of this study was to refine rural matching techniques in preparation for the 1990 Decennial Census .

The results contained in this paper show that computer matching can be accurate and effective in rural areas. The most effective refinement made to the computer matcher involved using the first letter of the last name as a blocking variable, as opposed to the 4 character SOUNDEX of the last name. Based on the results, we suggest using this blocking scheme for matching the 1990 census. The overall high match rate discussed above illustrates that a combination of computer and clerical matching can be effective in rural areas.

REFERENCES

Anolik, I. (1988), "The 1986 Rural Post-Enumeration Survey in East Central Mississippi," American Statistical Association Proceedings of the Section on Survey Research Methods, to appear.

Belin, T. R. (1988), "New Approaches in Computer Matching for Census Undercount," American Statistical Association Proceedings of the Section on Survey Research Methods, to appear.

Diffendal, G. (1988), "The 1986 Test of Adjustment Related Operations in Central Los Angeles County," Survey Methodology, 14, 71-86.

Fay, R.E., J. Passel, J. Robinson (1988), "The Coverage of Population and Housing in the 1980 Census," 1980 Census of Population and Housing Evaluation and Research Report, PHC80-E4. Bureau of the Census, Washington, D.C.: U.S. Department of Commerce.

Fellegi, I.P. and Sunter, A.B. (1969), "A Theory for Record Linkage," Journal of the American Statistical Association, 64, 1183-1210.

Hogan, H. and Wolter K. (1988), "Measuring Accuracy in a Post-Enumeration Survey," Survey Methodology, 14, 99-116.

Jaro, M.A. (1989), ""Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," Journal of the American Statistical Association, 64, 1183-1210.

Winkler, W.E. (1989), "Using the EM Algorithm for Weight Computation in the Fellegi Sunter Model of Record Linkage," American Statistical Association Proceedings of the Section on Survey Research Methods, to appear.

Footnotes

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau

[1] Prelist areas are areas in which an independent canvassing operation is done in order to compile a mailing list. Commercial mailing lists cannot be used in these areas because of insufficient coverage or inadequate mailing addresses.

[2] List/Enumerate areas are areas in which a door-to-door personal visit is conducted. These are primarily the more rural areas of the country, where commercial mailing lists cannot be used, and where it is not feasible to prelist.

[3] A work unit consists of one interviewer's work in one block on one day.