# USING ADMINISTRATIVE LISTS TO ESTIMATE CENSUS OMISSIONS: AN EXAMPLE

Eugene P. Ericksen, Mathematica Policy Research and Temple University
Joseph B. Kadane, Carnegie - Mellon University

## 1. THE PROBLEM

Decennial census omissions are inevitable: for example, some people never receive the forms, others are reluctant to answer and return them, and still others have uncertain residencies and are not included in the forms filled out by any household. The extent of omissions has been measured many times by the United States Census Bureau (Siegel and Jones, 1980). These studies have shown consistently that Blacks are harder to count than Whites, that men are more likely to be missed than women, and that young adult Black males are especially likely to be missed.

The Bureau's evaluations of the 1980 Census (U.S. Bureau of the Census, 1983) have shown three additional results. They are:

- o Hispanics are almost as hard to count as Blacks.
- o Blacks and Hispanics living in central cities are especially hard to count.
- o The omission rate is much higher than the net undercount, because omissions are partially offset by erroneous enumerations.

Erroneous enumerations include duplications, counts of persons who died before or were born after census day, or fictitious counts fabricated by enumerators. They are distributed rather evenly across geographic areas, while omission rates are especially high in urban areas with minority concentrations. As a result, erroneous enumerations offset omissions to a much greater extent in some areas than in others.

The impossibility of preventing omissions is demonstrated by the enormous resources which were poured into the 1980 Census effort. The Bureau spent one billion dollars (U.S. Bureau of the Census, 1980:92), more than four times the amount spent on the 1970 Census. The 1980 Census can be regarded as an ultimate test of whether complete coverage can be attained by conventional procedures. The answer clearly is no. The most recently available data from the Bureau's Post-enumeration Program, PEP, (1983) show that between 10 and 20 percent of Blacks and Hispanics were omitted from the 1980 Census in large central cities. Even after adjustments for erroneous enumerations are made, the net undercounts for these cities remain large.

There are many legislative apportionments and resource allocations which depend on the assumption of an accurate census. Not to take an accurate census shortchanges some areas and favors others. Realizing this, an unprecedented number of states and cities filed lawsuits against the Bureau in 1980. Because of the impossibility of counting everyone, and because of the political significance of this issue, there is a clear need for a statistical adjustment procedure which will lessen differential undercounting among areas.

Matching studies provide one basis for such a procedure. In Canada, matching studies referred to as "reverse record checks" appear capable of producing accurate estimates of omission rates for provinces (Fellegi, 1980). In the Canadian procedure, an alternative population list is compiled from lists of new births, new immigrants, people counted in the last census, and people found to have been missed in the last census. The list is sampled and matched against census records to determine omission rates. In the 1976 Census, of the 33,000 persons included in the sample, a match/nonmatch status was determined 95.7 percent of the time. Provincial omission rates varied from 0.38 to 3.13 percent , and these were greater than 2.5 times their standard errors in all but Prince Edward Island. There the omission rate was low, 0.38 percent, and the population small, 118,230.

In 1970, the U.S. Census Bureau carried out two matching studies. A study known as the "Medicare Record Check" (U.S. Bureau of the Census, 1973) used the list of Medicare recipients to represent the population aged 65 and over. Of the approximately 8,000 persons included in the sample, a match/nonmatch status was determined 96.5 per cent of the time, and precise estimates of omission rates were determined for Black and nonBlack males and females on national and regional bases. To estimate omission rates for the full population, the Bureau used the Current Population Survey, and has done this again for 1980 as part of the Postenumeration Program evaluating the Census. Research on the feasibility of basing match studies on the CPS continues at the Bureau. In addition to using the PEP, it is important to consider the use of matching studies based on alternative population lists. There are two reasons for this. One is that such studies might be better than those based on the CPS, and the other is that alternative population lists could become the basis for taking a more complete and efficient census (Alvey and Scheuren, 1982).

In the United States, it is not possible to compile a population list made up of nonoverlapping components as the Canadians have done. A reliable list of immigrants, documented or undocumented, does not exist. Moreover, with censuses spread ten years apart (five years in Canada), it is difficult to maintain a sample frame of persons not counted in the preceding census. We therefore consider it necessary to develop procedures based on sampling overlapping lists. An efficient method for removing duplicates without matching entire lists has been developed by Kadane and Lehoczky (1976).

The lists should be of two types: (1) representative, and (2) focused on the "hard-to-count". Lists of taxpayers, drivers, voters, and income tax exemptions will cover most people. Special lists of the hard-to-count can be obtained from sources such as rosters of welfare recipients, central city schoolchildren, people sending money orders overseas, or patients admitted to public hospitals.

Use of the alternative sample list may also aid matching. People can be located in the field through tracing procedures. Once located, all the necessary demographic information can be obtained to determine inclusion in the census. In addition the person can be asked if (s)he thinks (s)he was counted.

Our paper presents the results obtained from a 1980 Census match study based on a sample of overlapping lists in New York City. It was designed to estimate omission rates, which comprise only part of the undercount picture. However, because omission rates tend to vary considerably across areas while erroneous enumerations are more evenly distributed, and because estimates of erroneous enumeration rates can be based on samples taken from the census itself, we consider the accurate estimation of omission rates to be more important and technically more difficult.

## 2. SELECTING THE SAMPLE OF THE ALTERNATIVE POPULATION LIST

The City of New York has been a plaintiff in one of the lawsuits filed against the Census Bureau. The suit, originally filed in September, 1980, is still being litigated. As part of this litigation, the City constructed an alternative population list from local records. The Bureau was ordered to match a sample of this list against the census to estimate the omission rate for New York City.

The City compiled the alternative population list from ten local lists, including:

- o Consolidated Edison electricity billpayers
- o Babies born during the period immediately preceding Census Day
- o People who had died just after Census Day
- o New York City public schoolchildren
- o Persons arraigned in city courts
- o Students at the City University of New York
- o Persons included in the "Medicaid Eligibility File," primarily welfare recipients, and aged and disabled recipients of social security benefits.
- o Licensed drivers
- o Registered voters
- o Recipients of unemployment benefits

The City did not have access to nationally compiled lists such as those of taxpayers, food stamp recipients, or people listed as exemptions on income tax forms. This increased the number of lists necessary to substantially cover the City's population. Because there were so many lists, there was overlap, and a sampling procedure assuring that each person would have only one chance of selection was needed.

Sampling proceeded in three stages. First, an 8 percent sample of enumeration districts was selected. Next, samples of each of the lists were selected. Consistent with principles of optimal allocation, sampling rates were proportional to the square roots of the expected omission rates. Finally, the samples were matched against prior lists and duplicates were eliminated. The matching procedure worked as follows:

(1) The lists were numbered from 1 to 10, and used in this order.
(2) The sample from the first list was guaranteed inclusion in the final sample.
(3) The sample from the second list was checked against the entire first list, with any person included in both removed from the sample.
(4) For remaining lists, the samples were checked against all preceding lists, with duplicates removed.

This procedure substantially reduced the number of matching attempts. For sample selections of the kth list, it was only necessary to match against the preceding (k - 1) populations. As a result, the sample from the kth list included only those sampled people not included on any preceding list (see Kadane and Lehoczky, 1976).

Inter-list matching occurred in two general ways. Each list member had a name and address, and these were matched first. To be considered a match, the first three letters of the first name and first five letters of the last name had to be the same, and the addresses had to match. This reduced the chances of preventing a match because of spelling errors. The efficiency of this matching was enhanced by the ability to limit checks to those addresses included in the 8 percent sample of enumeration districts.

The second match was of names, again on the "three-five" basis, and other characteristics, such as sex, age, or social security number. Tolerances of one year for age matches were permitted. Checks of the possibility that the digits of the social security numbers had been scrambled were also made. These rules reduced the chances of failing to match when a match should occur, but increased the chances that false matches would also occur. On balance, this was a conservative rule, and it is likely that some names were falsely called duplicates and eliminated.

The final combination of lists produced a sample of 16,500 persons. When the reciprocals of their selection probabilities

were summed, the sample was found to represent a population of 6.2 million after duplicates were removed. This represents a large proportion of the City's population of approximately 7 to 8 million people. The Bureau was ordered to compare the list to the census and determine:

(a) the number of persons on the list who were counted in New York City in the 1980 Census;
(b) the number of persons on the list who were not living in New York City on April 1, 1980; and
(c) the number of persons on the list remaining.

Where necessary, persons not currently living at the sample address were to be traced to the current address so the appropriate determination could be made.

## 3. RESULTS

The Census Bureau matched the sample of names against census records during the late summer and fall of 1982, and submitted its report in November (U.S. Bureau of the Census, 1982a). The first step in the Bureau procedure was to consider a person as "matched" if (s)he could be found on the census form filled out at the sample address. In cases where no match occurred, the name was given to field workers who attempted to locate the person at the sample address. In some cases of movers, forwarding information was obtained from building superintendents, the Post Office, neighbors or people currently living at the address, and some tracing was done. A second group of people was found in the field, and for these a determination of whether or not they had been counted in 1980 was made. There was no determination of match status in the remainder of cases. It is likely that a substantial number of these were living at the listed address in April, 1980, were omitted from the census, and had since moved or died.

We have assigned weights to each of the sample cases equal to the reciprocals of the probabilities of selection. Having classified the cases into the three categories just described, their weighted distribution is:

| | |
|---|---|
| Located in New York City | 5.17 million |
| Counted in the census | 4.75 million |
| Not counted in the census | 0.42 million |
| No determination made | 1.03 million |
| Total | 6.20 million |

Of the estimated 5.17 million people included in the list and found to have been living in New York in April, 1980, 8.1 percent were classified as census omissions. This must be regarded as a lower bound of the list population omission rate.

There are two grounds for this statement. One is that the Bureau counted census imputations as matches. Imputations occurred in 1980 when the Bureau, after repeated attempts, had been unable to establish whether anyone, and if so, who, lived at an address. It eventually created household and person records by means of a computerized linking procedure. In the match study, there was no direct evidence that the persons on the sample list had actually lived at the addresses in question in 1980, but the Bureau called them "matches" anyway. Bureau practice in the Postenumeration Program was to count such a case as an omission, and then to subtract erroneous enumerations and imputations from omissions to obtain the net undercount (Cowan and Bettin, 1982:18). Approximately 0.9 percent of the matches were of persons at whose addresses an imputation was made. Most of these would have been counted as omissions in PEP.

The second basis for the statement is that many of the people for whom no determination was made were probably living in New York City in 1980 at the address supplied by the City. The bias of the Bureau's procedures is demonstrated by considering people living at listed addresses who moved between April, 1980 and August, 1982. If such a person were counted, the Bureau simply matched the person to a census form. If the person were not counted, the Bureau, hindered by limitations of time and resources, was not able to mount a comprehensive tracing effort. Such uncounted cases were typically classified as "no determination made" instead of as omissions. Lacking more complete tracing, the Bureau's procedure underestimates the omission rate. Of the "undetermined" cases, the Bureau reported that 35 per cent were cases where current occupants at the address had never heard of the person in question and 32 per cent were cases where the sample person had moved, but the current address was not reported. These are the subcategories where an above average omission rate is likely to have occurred. Bureau experience on PEP (Cowan and Bettin, 1982:12) suggests that movers were more likely than nonmovers to have been omitted. With these issues considered, it is likely that the correctly calculated omission rate was above 10 per-cent.

## 4. ARE THE ESTIMATES SENSIBLE?

The sensibility of the estimates can be evaluated in four ways, assessing (1) the proportion of cases where a matching decision could be made, (2) the consistency with other estimates, (3) the correlates of the omission rates, and (4) rules for dealing with missing data. The New York City match study results appear to be sensible on the first three criteria, not on the fourth. Better rules for dealing with missing data are needed. We first note that although matching took place 28 to 30 months after Census Day, and field tracing was limited, 83 per cent of cases were found to be living in New York City and a matching decision was made. When people found to have died or moved out of New York City are added to the "determined" group, the proportion of cases for which no determination was made is reduced to less than 15 per cent. The determined cases

represent over 5 million people, a substantial proportion of the list population, of the counted population of 7.1 million, and even of the full population which is somewhat larger.

Secondly, the estimates can be compared with the five omission rates made most recently available by the Bureau for New York City as part of PEP (U.S. Bureau of the Census, 1983). The estimates vary, because of different assumptions made for treating missing data, from 6.9 to 11.6 percent . The two estimates making the most sensible assumptions about missing data are 11.4 and 11.6 percent , not greatly different from the omission rate to be expected from the match study had more complete tracing efforts been made.

Thirdly, we have evaluated the geographic distribution of omissions. We first computed separate estimates of omission rates for each of the 20 "District Office Areas" in New York City, and then correlated these with mailback rates reported by the Census Bureau (Table 1). The mailback rates are the proportions of occupied households returning completed census forms through the mail. The most difficult census taking areas are those where mailback rates are low. In New York City, these areas have large Black and Hispanic populations. The overall mailback rate for New York City was 74 percent , well below the national average of 83 percent (U.S. Bureau of the Census, 1982b).

The mailback and omission rates are highly correlated, with the observed r = -.683 being dampened by the fact that the local omission rates include a random sampling component which could not be systematically related to mailback rates. The actual relationship is even stronger than this correlation indicates. In middle income areas like Staten Island and parts of Queens, people were relatively easy to count, and the omission rates were about five percent . In poorer areas like the South Bronx and Bedford-Stuyvesant (North Central Brooklyn) omission rates were consistently above 10 percent, with the high rate of 17.9 percent obtained in Bedford-Stuyvesant. Collapsing the District Offices into four groups and thus dampening the effects of sampling errors, the following results were obtained:

| Mailback Rates | Aggregated Omission rate (%) | District Offices |
|---|---|---|
| Below 70.0 | 13.4 | 7 |
| 70.0 to 74.9 | 9.1 | 5 |
| 75.0 to 79.9 | 6.5 | 6 |
| 80.0 and over | 6.4 | 2 |

The least sensible aspect of the matching study was the treatment of missing data. For reasons already discussed, it is likely that the omission rate among "undetermined" cases was greater than the reported rate of 8.1 percent. In order to make the best use of the alternative list methodology, a reasonable procedure for dealing with missing data is needed.

## 5. EVALUATION OF METHODOLOGY

The objective of the matching study, of course, was to estimate the omission rate for New York City, not simply for the persons on the alternative population list. There are two important steps necessary for improving the estimates. The first, and most important, is to reduce the proportion of cases in the sample where the match status was undetermined. The best strategy for this would have been a more complete tracing effort by the Census Bureau. Failing this, complete tracing of a sample of cases where the match status is undetermined would be useful. If that is impossible, some kind of inferential modelling is necessary. A crude sort of model is hot-deck imputation, a standard Census Bureau procedure used in the Decennial Census, the Postenumeration Program and surveys like the CPS. In hot-deck imputation, a neighboring case on the tape with determined match status is found having characteristics similar to the undetermined case, and then the match status of the determined case is attributed to the undetermined case. Variables that might be used to link cases should include the list from which the case was selected, the geographic location of the address and various demographic characteristics.

The principal failing of hot-deck imputation is that analysis using it pretends that the case with undetermined match has a known matching status. Thus information generated by a computer routine is treated as if it were information about the empirical world. A less crude model is to treat the unknown match status of an undetermined case as a random variable. Probabilities should be given for the various possible values of the undetermined cases. To give a hypothetical example, after the linked cases were assessed, we might estimate a probability of 50 percent for the status "matched," 30 percent for the status "omitted" and 20 percent for the status, "not in New York City in April, 1980." These probabilities would then be used to compute the distribution (or at least the asymptotic momemts) of whatever random variables are of interest. The variances report how much uncertainty is caused by the failure of the Bureau to complete the tracing of cases, and would add to uncertainty due to sampling.

Once the problem of missing data for list members was solved, the issue of relating the omission rate for the list population to the omission rate for the nonlist population must be dealt with. This is probably a less difficult problem. The objective in constructing the alternative population list is not so much to include everyone, but to create a list which is representative of the total population with respect to the likelihood of omission. The alternative list population should be compared to the nonlist

population with respect to variables known to be related to the likelihood of omission. Afterwards, list members selected from harder or easier to count lists could be reweighted to increase the representativeness of the list. This is the strategy followed by the Census Bureau in weighting the Current Population Survey to an independent demographic estimate when computing omission rates on PEP (Cowan and Bettin, 1982:29).

For future applications of the alternative list population strategy, it maybe more efficient to work with a small number of representative lists. On a national basis, the lists of income tax exemptions and social security recipients, advocated by Alvey and Scheuren (1982) could be used for the general population. Lists of the "hard-to-count" should then be searched for on a local basis, since the coverage of welfare and other lists might vary from place to place. Alternatively, the Current Population Survey could be used to represent the general population, and the local lists used for the hard-to-count. The CPS would be incorporated into the alternative population list methodology by calling the theoretical "CPS covered" population the last list, and checking sample members against all preceding lists.

REFERENCES


Alvey, Wendy, and Fritz Scheuren, "Background for an Administrative Record Census," Proceedings of the Social Statistics Section of the Annual Meetings of the American Statistical Association. 1982.

Cowan, Charles D., and Paul J. Bettin, "Estimates and Missing Data Problems in the Post Enumeration Program," unpublished Census Bureau memorandum. 1982

Fellegi, Ivan, "Should the Census Count be Adjusted for Allocation Purposes: Equity Considerations," Proceedings of the 1980 Conference on Census Undercount, Washington, Government Printing Office, 1980.

Kadane, Joseph B., and John P. Lehoczky, "Random Juror Selection from Multiple Lists," Operations Research, Vol. 24, 1976.

Siegel, Jacob and Charles D. Jones, "The Census Bureau Experience and Plans," Proceedings of 1980 Conference on Census Undercount, Washington, Government Printing Office, 1980.

United States Bureau of the Census, "The Medicare Record Check: An Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census," Census of Population and Housing: 1970, Evaluation and Research Program PHC(E)-7, Washington, Government Printing Office, 1973.

United States Bureau of the Census, Census '80: Continuing the Factfinder Tradition, Washington, Government Printing Office, 1980.

United States Bureau of the Census, "Report of the United States Bureau of the Census in Response to the Order of the Court Ordered August 6, 1982," report submitted to the Federal Court of the Sourthen District of New York, November, 1982(a).

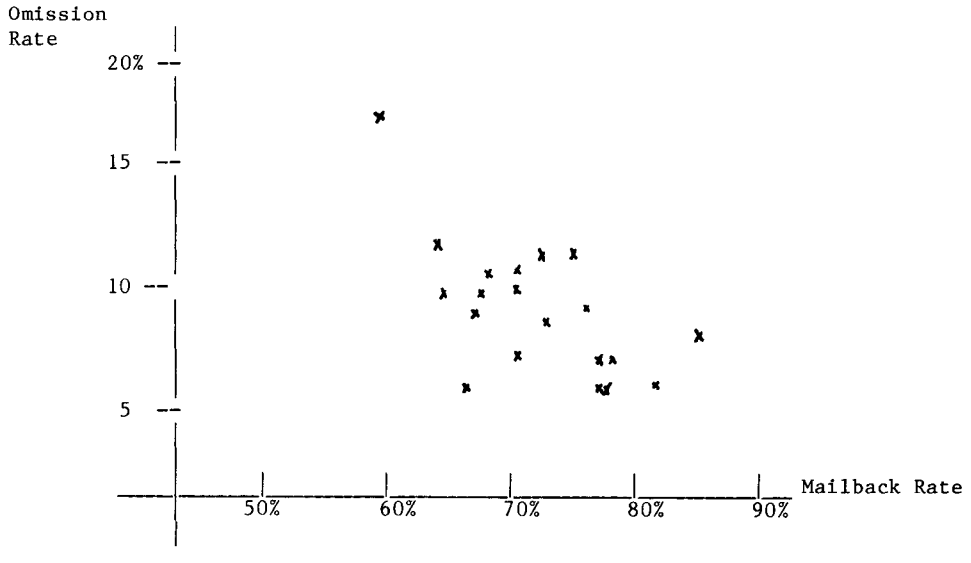
United States Bureau of the Census, "Decentralized Office Worksheet": unpublished Census Bureau memorandum submitted to the Federal Court of the Southern District of New York, December, 1982(b).

United States Bureau of the Census, unpublished tabulations submitted to the Federal Court of the Southern District of New York, January, 1983.

TABLE 1

Relationship Between Mailback Rate and Omission Rate

| District Office | Area | Mailback Rate | Omission Rate |
|---|---|---|---|
| 2252 | N. Central Brooklyn | 59.9% | 17.9% |
| 2254 | Central Brooklyn | 62.8 | 12.0 |
| 2241 | N. E. Manhattan | 63.8 | 9.5 |
| 2249 | N. W. Queens | 66.8 | 5.7 |
| 2248 | S. W. Bronx | 67.4 | 8.9 |
| 2244 | W. Manhattan | 67.9 | 9.6 |
| 2250 | S. E. Queens | 68.4 | 10.5 |
| 2256 | S. E. Brooklyn | 70.1 | 9.9 |
| 2245 | S. Bronx | 70.2 | 10.4 |
| 2247 | N. W. Bronx | 70.3 | 6.6 |
| 2253 | N. E. Brooklyn | 71.5 | 10.9 |
| 2251 | N. W. Brooklyn | 73.3 | 7.8 |
| 2240 | N. Manhattan | 76.2 | 10.9 |
| 2243 | S. Manhattan | 76.9 | 8.3 |
| 2255 | S. Brooklyn | 78.0 | 6.2 |
| 2242 | E. Manhattan | 78.2 | 5.1 |
| 2201 | S. W. Brooklyn, Staten Island | 78.8 | 5.1 |
| 2202 | N. E. Queens | 79.0 | 6.0 |
| 2203 | W. Queens | 82.5 | 5.4 |
| 2246 | E. Bronx | 85.0 | 8.2 |
| | Total City | 74.2 | 8.1 |




NOTE: Mailback rates were obtained from "Decentralized Office Worksheets," unpublished memorandum, U.S. Bureau of the Census, December 1982.