Journal of Official Statistics
Vol. 2, No. 4, 1986, pp. 397–414
© Statistics Sweden

Using Administrative Lists to Estimate Census Omissions

Eugene P. Ericksen[1] *and Joseph B. Kadane*[2]

Abstract: We present a method for estimating omission rates from censuses. Our method is based on the merger of administrative lists, sampling from these lists, and matching against census rolls. We describe the method, present the results of a test in New York City, U.S.A. in 1980, and evaluate the results. We compare our proposed method to other procedures for estimating omission rates.

Key words: Administrative lists; censuses; matching; omission rates.

1. Introduction

In the United States, public policy makers rely on census data for very important purposes. For political apportionment, legislative districts must have practically equal populations; the courts have mandated very narrow tolerances (*Karcher v. Daggett* (1983)). For local government financing, census counts are the basis for allocating billions of dollars of revenue-sharing and other funds. With so much at stake, it is easy to understand why cities and states are concerned about the occurrence of census errors. Studies of the American undercount have shown that Blacks are missed at a much greater rate than non-Blacks, usually at ratios between 3 to 1 and 5 to 1. Recent data support the common sense conjecture that Hispanics are missed at rates similar to Blacks and that undercount rates are particularly high in large central cities (National Research Council (1985), Ericksen and Kadane (1985, 1987)). The circumstances of high stakes and differential undercounts have created turmoil, and in 1980 over 50 cities and states sued the U.S. Census Bureau to have the census counts adjusted to compensate for undercounting. Many statisticians and demographers participated on both sides of the lawsuits and the question of adjusting for the undercount has become an important controversy among American statisticians and demographers (Ericksen and Kadane (1985), Freedman and Navidi (1986)).

Discussions seem to coalesce around two issues. First of all, there is growing acceptance of the inevitability of the undercount. As stated in the final report of a panel of the National Academy of Sciences especially appointed to review the data of the 1980 Census and to make recommendations concerning the 1990 Census: "The evidence is overwhelming that no counting process, however

[1] Professor, Department of Sociology, Temple University, Philadelphia 025–23, PA 19122, U.S.A.
[2] Professor, Department of Statistics and of Social and Decision Sciences, Carnegie-Mellon University, Pittsburgh, PA 15213, U.S.A.

Acknowledgement: The authors would like to acknowledge the efforts of the following New York City government officials, without whom the study could not have been done: Wendy Smyth, Evelyn Mann, Robert Amsterdam, and David Berger.

diligent, will in fact enumerate everyone (National Research Council (1985, p.17))." Moreover, efforts to count everyone become increasingly costly as special efforts are made to find the hard to count (Keyfitz (1979)). In 1980, out of a total budget of $1 billion, the Census Bureau spent $200 million on coverage improvement programs and still omitted millions of persons (National Research Council (1985, pp. 227–231)). In addition, procedures used to improve coverage caused other erroneous enumerations to occur (Ericksen and Kadane (1985), National Research Council (1985, pp. 183, 201–202)). For example, some housing units were listed twice as the result of rigorous field checks made to assure that no housing units were missed. A recent Census Bureau report concluded that the housing unit duplication rate in the 1980 Census was 0.86 per cent (U.S. Bureau of the Census (1985)). This means that about two million people were double-counted because they lived in houses that were counted twice. Census-takers cause erroneous enumerations when they fabricate people, or double-count persons with more than one residence. Census Bureau statisticians Cowan and Fay (1984) estimate that six million erroneous enumerations were included in the 1980 Census count of 226.5 million.

The undercount is the net between omissions and erroneous enumerations and is estimated by comparing the census count to a demographic estimate of the national total. The 1980 estimated net undercount for Blacks was between 4 and 6 per cent and for non-Blacks was a figure near zero (U.S. Bureau of the Census (1982a)). The Census Bureau calculated alternative, survey-based estimates in a project known as the Post Enumeration Program (PEP). The survey estimates confirmed the Black/non-Black differential found by demographic procedures, and also showed that (1) Hispanic undercounts were comparable to those of Blacks and (2) undercount rates were higher in the central cities. The Census Bureau made a valiant, but expensive and ultimately unsuccessful effort to count everyone in 1980. It seems reasonable to conclude that this cannot be done in a cost effective manner.

The second issue is the feasibility of adjustment. Some (Trussell (1981), Freedman and Navidi (1986)) argue that no valid adjustment method exists, while others (Tukey (1983), Fisher (1984), Ericksen and Kadane (1985)) argue the opposite. The controversy over census adjustment in the United States continues. In general, the argument does not concern the need for an adjustment, but whether the proposed adjustment methods will produce sufficient improvement over the unadjusted counts.

Our objective in this paper is to present and evaluate a method for estimating one part of the undercount – omission rates. In brief, our method involves constructing a sample from a compendium of administrative lists, matching this sample against census records, and estimating omission rates from the results. Afterwards the estimates can be combined with separate estimates of erroneous enumeration rates to produce a final estimate of the net undercount. First we describe the Census Bureau's method for estimating omissions in the PEP. Then we review past research on the uses of administrative records and other lists in censuses. We describe our method and its test, and we close the paper with a general discussion of our method's strengths and weaknesses.

2. The Census Bureau's Method for Estimating Omission Rates in 1980

The Census Bureau's method was not based on a list of persons, but on the Current Population Survey (CPS), which is, theoretically, representative of the civilian noninstitutionalized population. It is a monthly survey primar-

ily used for estimating the unemployment rate and it is based on a sample of households and noninstitutional group quarters (see U.S. Bureau of the Census (1978) for a description). The PEP included all persons in the April and August, 1980 editions of the CPS and matched them against the census rolls. A CPS sample person found in the census was classified as "matched," but a sample member searched for and not found was classified as "unresolved" and designated for a field check. The Census Bureau sent an interviewer to the field to find the person and obtain the correct April, 1980 address. Census returns were then rechecked to confirm that the person had in fact been missed and, if so, the person was reclassified as an "omission." In other cases, field checking uncovered information indicating that the sample person had been counted and (s)he was reclassified as a match. Remaining cases, about 4 per cent of the total, were assigned a status of "unresolved," usually because the CPS data were of poor quality, the person had moved to an unknown destination by the time field checking took place in the fall of 1980, or the sample person was found to have been fabricated by the CPS interviewer. The Bureau devised two major strategies for such cases.

One strategy considered such cases to have been omitted at the same rate as the rest of the sample. Because the large majority of the sample had been matched, the strategy in effect led to the assumption that most unresolved cases were matches and produced a lower estimated omission rate. The second strategy linked such cases to the other previously unresolved cases that had been resolved in the field, and the count/omission status of the latter was imputed to the former. Since approximately half of the cases resolved in the field were found to have been omissions, this increased the omission rate. For the April data, the omission rates estimated under these two assumptions were 3.66 and 5.55 per cent, a difference of 1.89 per cent.

Matching problems increase the difficulty of using the CPS as the source of matching in a Post Enumeration Program. The fact that the CPS does not reach everyone is a second problem. When the U.S. Bureau of the Census (1978) compared population estimates based on the CPS sample with the 1970 counts, they found that about 10 per cent of Blacks and 3 per cent of non-Blacks were missed. Most likely, the persons missed by the CPS are the same types of people missed by the census, making omission rate estimates based on comparing the CPS to the census conservative.

We hoped that using the lists of persons living in New York City would solve some of these problems. We included lists of persons who were more likely to be hard-to-count. In addition, because the list information included name and address and usually some demographic data, we hoped that this information would be of higher quality than the CPS data and would better facilitate matching. Our approach was similar to that used by the Census Bureau in 1970 in the Medicare Record Check (U.S. Bureau of the Census (1973)).

3. Earlier Attempts to Use Administrative Lists to Improve Censuses

There are two ways in which administrative lists have been used in censuses. One is to provide an administrative population list, sample it, and match the sample against the census to estimate an omission rate as was done in the Medicare Record Check. The other, more radical, procedure is to use an existing or to construct a new population register and to substitute this for the count. This has been done in Denmark (Redfern (1983)) and has been suggested for the United States (Alvey and Scheuren (1982)). Our approach is the first, but we will briefly discuss the radical alternative to provide additional understanding of some of the advantages and disadvantages of the method we propose.

In Denmark, the government based its census on a central population register, a list that should include all residents. This register was merged with other government files to obtain demographic, economic, and other characteristics. The Danish data files appear to be relatively complete, but important problems remain. One such problem is that the central register includes people who have emigrated. Another is that some of the other lists have out-of-date or otherwise erroneous data. To our knowledge, no one has studied the omission rate of the Danish census, so there is no evaluation of the completeness of the lists. Any undocumented aliens living in Denmark present an additional problem.

The lack of a central population register in the United States increases the difficulty of constructing an alternative population list. Alvey and Scheuren proposed constructing such a list by combining income tax, social security, and other files, an expensive and untested proposition. Other substantial problems exist. First, some people are not included in any of the lists being combined. Adults not paying taxes nor receiving social security benefits are excluded, so many poor people and undocumented aliens are left out. Second, some of the lists have poor address information, for example, some people use a post office box as the address on their income tax forms. Third, identifying information for the same person can vary between lists. When the lists are matched against each other, the duplication of a person will not be identified and (s)he will be included twice. This often happens when one list is older than another, and the person has moved in the meantime. However, this problem is avoided for the many lists including social security numbers.

These problems seem formidable, and to our knowledge no one has seriously considered using Alvey and Scheuren's suggestion. Our proposal modifies theirs. By adding lists with concentrations of the poor and hard-to-count, we hope to alleviate the first problem. By matching persons on characteristics other than addresses, we hope to alleviate the problem caused by changed addresses. By sampling we lessen the cost. By matching the sample against the census, and using field checks to verify a person's omission from the census, we hope to lessen the effects of incomplete and incorrect information on the list. As long as the alternative population list is representative both of hard and easier-to-count persons, we do not consider it crucial that the list be complete to evaluate the completeness of a census count. The match of an alternative list sample against the census can indicate the rate of omission and provide information on the types of persons missed and their addresses. This is an important, but less ambitious objective than that of compiling an alternative population register. Methods similar to our procedure have been tested in the United States and Canada.

The Medicare Record Check is one important example. The Medicare program in the United States provides health insurance for nearly all persons age 65 and over. In 1970, Census Bureau statisticians selected a sample from the list of eligible recipients to match against census returns and estimate their omission rate. Since eligibility for Medicare is nearly universal, the list of recipients provided a good population list of this group. Matching was successful, since 96.5 per cent of the sample was traced to a census day address, and the count/omission status was determined. The Census Bureau concluded that 4.9 per cent of all recipients, but 11.0 per cent of Blacks and recipients of other races were omitted from the 1970 Census (U.S. Bureau of the Census (1973)).

The Medicare Record Check was successful, but for a limited subgroup of the population. Again in 1970, the U.S. Bureau of the Census (1971) selected a sample of licensed male drivers aged 20 to 29 living in Washing-

ton, D.C. Comparing this group to census records, they found that 14 per cent were definitely missed by the census, but there was a large additional group for which a determination could not be made. One important problem concerned out-of-date addresses. The address used for obtaining the driver's license was not necessarily the one at which the sampled person lived at the time of the study and the Census Bureau found it hard to trace the sample. The conclusion to be drawn from these two studies is that administrative lists can be used to estimate omission rates for certain well-defined subpopulations, but substantial problems remain for other important groups.

In 1980, the Census Bureau matched a sample of 1979 tax filers to census records. For those persons where a match status could be determined, a 12.6 per cent omission rate was estimated (National Research Council (1985, p. 230)), but again, serious matching problems were caused by out-of-date addresses. In addition, many poor people, more likely than average to have been missed, did not file income tax forms and were not included in the study.

A more successful study has been made in Canada (Fellegi (1980)). The Canadians have studied the completeness of their recent censuses through the matching of a sample selected from a composite of four lists: (1) persons counted in the previous census, (2) persons found not to have been counted in the previous census, (3) births since the previous census, and (4) persons immigrating into Canada since the previous census. In 1976, a sample of 33 000 persons was selected from this composite, and successful determinations of match/nonmatch status were made 95.2 per cent of the time. The national omission rate was estimated to be 2.0 per cent, and provincial omission rates varied from 0.38 to 3.13 per cent. The Canadian study appears to have been more successful than the three American studies for two main reasons. One is that their composite of four lists covers the population better than any list available to the Americans, and the second is that their address information is more up-to-date. This is partly because most people were counted in the previous census which was taken only five years earlier, rather than at ten year intervals as in the United States. Given the high mobility rates in the United States, most experts think that ten-year-old addresses are not good enough for efficient tracing and matching. The problems of incomplete lists and out-of-date addresses appear to be the most serious obstacles to the use of lists for estimating census omission rates in the United States.

4. The Administrative List Study in New York City

New York City, where 44 per cent of the counted population was Black or Hispanic, was one of the cities that sued the Census Bureau in 1980. One of the issues of this lawsuit was whether or not the estimates of omissions provided by the match of the Current Population Survey to the census are sufficiently accurate to demonstrate that a large undercount existed in New York City and ultimately to adjust the census. In an attempt to provide an alternative and perhaps better estimate of the New York City undercount, the city proposed what has come to be known as the "megamatch study": that a compendium of administrative lists of New Yorkers be compiled, sampled, and matched against the local census rolls to estimate the local omission rate. The judge hearing the case ordered that the study take place. Accordingly, New York City provided the sample of names drawn from the lists and the Census Bureau matched this against the rolls and estimated an omission rate. Results are described by Bailar (1983, pp. 68–73) and Ericksen (1983, pp. 94–

101). We now describe the steps of this procedure:

First, the city compiled an alternative population list from the following ten local lists:

Persons included in the "Medicaid Eligibility File," primarily welfare recipients and aged and disabled recipients of social security benefits,

Billpayers to the local electric company, Consolidated Edison,

Babies born during the period immediately preceding census day,

People who had died just after census day,

Registered voters,

Licensed automobile drivers,

Persons arraigned in city courts,

Students at the City University of New York,

Recipients of unemployment benefits, and

New York City public school children.

The city did not have access to nationally compiled lists such as persons listed on income tax forms or persons with social security numbers. None of the locally available lists included even half the city's population, so we needed many lists to cover the majority of New Yorkers. There was a substantial amount of duplication – the total number of names on all ten lists was around 20 million, much greater than the city population of 7 to 8 million. We needed a method to eliminate duplicates, especially since many lists lacked social security numbers, unique identifiers of persons. To solve the problems of sampling and identifying duplicates, we adopted a sampling and matching strategy similar to that used by Kadane and Lehoczky (1976) to compile juror lists.

The sampling strategy included three steps. First, we selected an eight per cent sample of enumeration districts – small geographic areas defined by the Census Bureau to make enumerator assignments – and we created "sublists" including just those persons on each list who lived in one of the sampled enumeration districts. Next, we selected samples within each of the ten sublists. Consistent with the principles of optimal allocation, we set sampling rates in proportion to the square roots of the expected omission rates. For example, we expected unemployed persons to be missed at a higher rate than electricity billpayers and they were sampled at a higher rate. This was in approximate proportion to $\sqrt{(pq)}$, where p is the omission rate and $q = (1-p)$, close to 1.0.

Finally, we matched the samples, rather than the full populations, against prior lists and eliminated duplicates. This was a key step that created a very large reduction in the cost and time of completion of sampling. 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 matched against the entire first list, with any person included in both removed from the sample, and

(4) for the remaining lists, the samples were checked against all preceding lists, with duplicates removed.

While the sample from the second list was checked against only the one preceding list, the sample from the tenth list had to be checked against all nine of the lists which preceded it. In general, a person selected from the kth list was checked against the preceding $(k-1)$ lists and the final sample from the kth list included only those persons not included in any preceding list. The procedure provided an efficient alternative to Alvey and Scheuren's proposal of compiling an alternative population register. Each person included in one or more lists has the same chance of selection as any other on the first list in which (s)he is included. By sampling first and then matching, we do not have to compile the complete list of millions of people. Of course, both our procedure and the Alvey-Scheuren procedure rely on correct matching.

Inter-list matching occurred in two ways, one involved addresses and the other involved

other characteristics. This two-step matching strategy was intended to lessen the problem of lists that were compiled at different times and on which the same person could be listed twice with different addresses. The first method involved the name and address, available for each person on each list. When we matched a sample name against the names on the previous list, three things had to be true before we considered a match to have occurred: (1) the first three letters of the first name had to be the same, (2) the first five letters of the last name had to be the same, and (3) the addresses had to be the same. This strategy reduced the chances of preventing a correct match due to spelling errors. On the other hand, it permitted some false matching to occur. For example, a husband and wife named Bernard and Bernice Greenberg found on separate lists would be considered a match. Because our method was new, we chose to adopt a conservative strategy which probably had the effect of making the final sample smaller than it would have been with perfect matching.

The second match was of names, again on the "three-five" basis, and characteristics other than address, such as sex, age, or social security number, depending on the variables available on the two lists being matched. This allowed the matching of persons with different addresses on different lists. Tolerances of one year for age matches were permitted. Checks of the possibility that social security numbers had been scrambled were made by matching all permutations of the numbers given on each list against each other. This is a technique commonly used by government agencies checking welfare lists against tax records to prevent fraud. Again, these rules reduced the chances of failing to match in a situation where a match should occur, but increased the chances that false matches would occur.

On the other hand, we limited matches to those addresses in the eight per cent sample of enumeration districts. If a sampled person was listed on a preceding list at an address outside the eight per cent sample, the match would not have been identified. This created possibilities of failing to match the same person with different addresses on two lists and therefore of increasing in an undesirable way the chance of including the person in the sample. To reduce the likelihood of such errors, we eliminated from the sample persons who had been on a list for more than two years. For example, only those licensed drivers who had obtained new or renewed old licenses in the last two years were included in the study, eliminating a substantial number of persons with older licenses. On balance, our matching rules probably caused more false matches than false nonmatches.

The final combination of lists with duplicates removed produced a sample of 16 536 persons. Weighting sample members by the reciprocals of selection probabilities, our estimate of the number of persons represented by the sample is 6.2 million. This is a substantial proportion of the 7 to 8 million persons living in New York City in 1980.

The Census Bureau was ordered by the court to compare the sample list to census records, and to 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.

The Census Bureau procedure included two steps. First, the Bureau matched the sample against census records. Persons found there were considered to be matched. Remaining persons were traced to the current address wherever possible and interviewed to determine the correct census day address and consequently whether or not that person had been counted. The difficulty occurring because

tracing took place in September and October of 1982, over two years after census day, was a result of the litigation process and would not normally be a problem.

5. Results

The Census Bureau determined a match or omission status for a weighted sample sum of 5.17 million people (Table 1), 83.4 per cent of the total. The remaining 16.6 per cent were unclassified for one of two reasons: (1) the Bureau determined that the person was not living in New York City on census day, or (2) the Bureau was not able to determine the April 1, 1980 address. About half the unclassified cases fell into each of these categories. The Census Bureau was therefore able to determine in over 90 per cent of cases whether a sample person was counted, omitted, or not living in New York City on April 1, 1980. Given the passage of 2 1/2 years since census day, this is an impressive result which would surely be improved upon if the time between enumeration and field tracing were reduced.

Table 1. Results of Census Bureau Matching Efforts, Weighted Data

Disposition of case	Number of persons	
Counted by the census in New York City	4 750 000	
Omitted from the census count in New York City	420 000	
Count/omission status not determined	1 030 000	
Not living in New York City on census day*		510 000
Address on census day not determined*		520 000
Total	6 200 000	

* Detailed results for weighted data were not provided by the Census Bureau. These estimates assume that the distribution of outcomes within the category of cases where the count/omission status was not determined is the same for weighted as unweighted data.

Turning now to the 5.17 million names where the match status was determined, a weighted total of 420 000, 8.12 per cent, was found to have been omitted from the census. This result must be evaluated in the face of at least two sources of uncertainty: (1) sampling errors, and (2) uncertainties in matching decisions. We have estimated the sampling error to be approximately one-half of one per cent, leaving the uncertainty of matching to be the major source of uncertainty.

We intended the addresses from the administrative lists to refer to a date near census day, 1980 rather than the matching and tracing period in 1982. We hoped that this would facilitate matching against the addresses recorded in the census but realized that tracing the many people who had moved between 1980 and 1982 would be more difficult. The Bureau's policy was not to classify a person as an omission unless reliable information concerning the census day address, obtained from the person in question, was found through field tracing. This had the effect of making matching rules conservative. Movers found in the census rolls were classified as matches without a field check but movers not found in the census rolls could not be classified as omissions without the field check. Any sample person who had been omitted by the census but had since moved would not normally be classified as an omission unless the Census Bureau traced them to the new address. With the passage of time,

finding such people was not an easy task for the Census Bureau. Consequently, movers would usually be counted as matches if they were truly counted, but were less likely to be counted as omissions if they were truly omissions. Instead they were often included among the 1.03 million unclassified persons and this had the effect of lowering the estimated omission rate.

We must be careful when we draw conclusions about the entire population of New York City on the basis of this study. There are perhaps 2 to 2.5 million people not included among the lists for whom the omission rates could be either higher or lower. If, however, we assume that the omission rate among the total population was 8.12 per cent, then the published count of 7.072 million persons leads us to conclude that 625 000 New Yorkers were omitted by the census.

The estimate of 8.12 per cent provides important information about the undercount in New York City and we are confident that improved results could be obtained in future applications when there would be a shorter period between census day and the time of matching. In the meantime, we can evaluate our procedure in four other ways: (1) the classification rules, (2) the quality of the lists, (3) the relationship between difficulties in taking the census and difficulties in the matching process, and (4) comparing the results of this study with those of the Post Enumeration Program in New York City.

6. Evaluation of the Procedure: Classification Rules

The Census Bureau produced two tabulations of results of the matching study. The first, already discussed above, separated results by list and permitted calculation of the weighted results for the major categories of matching status we have discussed above. A second, more detailed tabulation (Table 2 on the following page), was also provided but without the information needed to calculate weighted results (U.S. Bureau of the Census (1982b)). It is not likely, though, that ignoring these weights changes results greatly.

The more detailed results help us to understand some of the difficulties of matching and the types of persons who were omitted. For example, there were five subcategories of census omissions. In only 89 cases, well under ten per cent of all omissions, were entire households missed. Our list based approach, not sorted by household, appears to have been particularly good at finding missed persons within households otherwise counted but did not identify many people living in households where everyone was missed. There were many other cases where the census-takers' work was simply incorrect. For example, there were 650 cases where another household had been counted at the address where a sample member reported that (s)he had lived on census day. This family could have moved in after census day, the enumerator could have gotten addresses mixed up, or the counted family could have been fabricated by the census-taker. Some of these are erroneous enumerations, and this demonstrates the necessity of taking erroneous enumerations into account when estimating the undercount.

Turning to the matches, we see that most of the matches occurred at the addresses shown in the New York City lists. Indeed, 11 792 of these, 71.3 per cent of the entire list total, matched at the address supplied by the city and 716 additional matches were found elsewhere in New York City. Finally, there were 112 cases of sample persons whose addresses were at households where all census person-records had been imputed, or created by computer. The imputations occurred in the following types of situation: after repeated attempts, census-takers could not determine whether a housing unit was occupied or vacant, or who might live in a housing unit

Table 2. Detailed Results of Census Bureau Matching Efforts, Unweighted Data

Disposition of case	Number of cases	Per cent of classified cases	Per cent of all cases
Counted by the census in New York City	12 620	91.7	76.3
Matched at address given by New York City	11 792		
Matched at a different address	684		
Name matched, address undetermined	32		
Matched at an imputed household*	112		
Omitted from the census count in New York City	1 146	8.3	6.9
Apparent miss, others counted at address	312		
Apparent missed housing unit and person	89		
Different household at address	650		
Duplication or imputation in separate unit in same building	61		
Moved out or died during census period	34		
Not living in New York City on census day	1 379	–	8.3
Moved in after or out before April 1	873		
Died before April 1	310		
Should not have been counted at this address	196		
Address on census day not determined	1 391	–	8.4
Non-residential sample address	270		
Sample person unknown	983		
Insufficient information	138		
Cases classified by Census Bureau	13 766	100.0	83.2
Total cases	16 536	–	100.0

* If the address of a sample person was at a household where all person-records in the census had been imputed, or created by computer, the Census Bureau considered the person to have been matched against one of the sample persons.

SOURCE: United States Bureau of the Census (1982 b).

determined to be occupied. In such cases, the Census Bureau had its computer impute an occupancy/vacancy status to the unit and/or artificially create persons to be living in those units imputed or determined to be "occupied." In the megamatch study, the Census Bureau classified the 112 cases as matches, a strategy with which we disagree. In the Post Enumeration Program, such a case was counted as an omission (Cowan and Bettin (1982, pp. 6, 18)), and we feel that this should properly be done here. Such a decision increases the estimated omission rate to 9.1 per cent (see also Bailar (1983, pp. 68–71)).

Next, we turn to the problem cases, where the Census Bureau could not make a matching decision. In about half of these, the Bureau learned that the person was not in fact living in New York City on census day, so about 8 per cent of our weighted total of 6.2 million should properly be removed from the study. In the other half of cases, the Census Bureau could not make a determination. It is often likely that the person had moved since the list was prepared, and the housing unit was destroyed or left vacant, or is now occupied by persons who know nothing of the listed person. Because these persons were not found on the

census rolls in New York City, it is likely that many had moved away and were counted elsewhere but that a substantial number were in fact omitted. There were 1 391 such cases and it seems reasonable to conclude that the omission rate was somewhat above ten per cent. This would be true if as few as one-third of them were omissions.

The estimated number of persons whose census day address was not determined (520 000) is greater than the number of omissions (420 000). This need not be a problem if a suitable imputation model is available, as in the case in the 1980 Census itself. For example, in New York State, income was imputed for 850 000 families, 19.1 per cent of the total, by linking missing data cases with families having similar characteristics and assuming that the incomes were the same. The number of families with income imputed was greater by far than the finally estimated number of families living in poverty, 480 000. The poverty statistic, needed to inform many important government policies, is considered to be suitable for use. In the same spirit, we advocate linking cases on variables related to the likelihood of omission, such as list membership and geographic location, to impute statuses of count/omission/not living in New York City. Such a strategy would produce an improved estimate of the omission rate.

7. Evaluation of Procedure: Quality of Lists

It is likely that the qualities that make a person hard to count would also make it hard to determine the true count/omission status. Such people might be more likely to live in housing units that are hard to find, they might be more likely to move, or they might be more likely to try to avoid census-takers, either on the original count or on the matching check. We were curious to learn whether such problems were concentrated in certain lists for two reasons: (1) so we would know which lists to avoid using in the future, and (2) so we would learn whether list problems or local population characteristics caused matching problems to be concentrated in certain areas.

In Table 3 (on the following page), we show, for each list and the entire sample, the rates by district office at which count/omission statuses were left undetermined. There were 20 district offices in New York City, each managing census-taking for an assigned section of the city. The local difficulty of taking the census is measured by the "mailback rate," or the proportion of occupied households who mailed their completed census questionnaires back to the Census Bureau. Nationally, 83 per cent of households mailed (returned) the forms, and this rate was achieved in just one of the twenty district offices in New York City. As we see in Table 3, mailback rates fell below 70 per cent in seven cases, reaching a minimum of 60 per cent in Office 2252, located in the Bedford-Stuyvesant section of Brooklyn. This area is well-known for being a difficult place to take a census.

The next four columns indicate the proportions of cases in four lists, identified at the end of the table, where the count/omission status was not determined by the Census Bureau. We note that the first list was actually a compendium, referred to as the "confidential" list, of seven individual lists merged to further protect the privacy of individual list members. The next column gives the proportion of cases in each district office where the count/omission status was not determined, and the final column gives the proportion which would have remained undetermined had the rates applying to each list for the entire city been applied to each district office.

We first note that 21 per cent of cases on the confidential list remained undetermined, a much higher proportion than that obtained on any other list. This is not surprising, since the

Table 3. Relationship Between Mailback Rates and Problems With Matching List Sample Against the Census

District office	Mailback rate (%)	Per cent of cases where count/omission status is undetermined				Merged list	Expected rate[2]
		List[1]					
		1	2	3	4		
2 246	85	17	12	18	9	16.0	16.6(+0.6)
2 203	82	17	7	12	6	13.2	16.9(−3.7)
2 201	79	13	7	13	2	11.8	16.8(−5.0)
2 202	79	19	14	9	6	14.0	17.0(−3.0)
2 242	78	24	20	15	14	19.9	17.5(+2.4)
2 255	78	15	8	12	0	12.0	17.0(−5.0)
2 243	77	23	28	14	11	20.5	16.9(+3.6)
2 240	76	27	14	26	16	23.0	16.8(+6.2)
2 251	73	17	12	10	13	13.9	16.7(−2.8)
2 253	72	24	6	17	9	14.8	15.9(−1.1)
2 245	70	22	8	12	5	13.8	16.0(−2.2)
2 247	70	28	10	21	7	19.6	16.1(+3.5)
2 256	70	22	10	11	4	15.2	16.7(−1.5)
2 244	68	16	30	14	6	16.9	17.5(−0.6)
2 250	68	23	9	11	4	15.5	16.3(−0.8)
2 248	67	35	13	28	9	22.9	16.0(+6.9)
2 249	67	22	14	13	5	16.1	16.5(−0.4)
2 241	64	28	21	24	6	23.3	16.7(+6.6)
2 254	63	21	16	16	10	17.1	15.9(+1.2)
2 252	60	35	24	16	16	25.6	15.7(+9.9)
Total city	74	21	14	14	7	16.6	

[1] Lists are defined as follows:
1. A compendium of seven lists, merged to assure confidentiality of list members, which included unemployed persons, CUNY students, voters, licensed drivers, births, deaths, and arraignees.
2. The Medicaid Eligibility File, primarily welfare and social security recipients.
3. Consolidated Edison (electric company) billpayers.
4. New York City public school students.

[2] Expected rate was calculated as $\Sigma(p_i r_i)$, where p_i is the proportion of the sample in the district office in list i and r_i is the total city rate at which the sample from list i could not have its match status classified by the Census Bureau. The difference between the actual and expected rates is shown in parentheses.

list included unemployed persons, arraignees, and CUNY students – highly mobile groups. Fourteen per cent of persons on the Medicaid Eligibility File and electric company billpayer lists, but only seven per cent of the New York City public school children were undetermined. We observe the general result in most district offices, with the confidential list usually having the highest proportion of undetermined cases (17 of 20), and the school children list having the lowest proportion (16 of 20, with one additional tie).

Next, we note the negative relationship between the mailback rate and the proportion of undetermined cases. We observe the general pattern for each of the lists, where census-taking problems were greater, matching was more difficult. The positive deviations between actual and expected rates in more difficult areas and negative deviations in less difficult areas indicate that local conditions independent of the quality of list information affected the ease of determining the count/omission status.

8. Evaluation of Procedure: Relationship Between Census-Taking Difficulties and Omission Rate

Next we ask whether the expected positive relationship between census-taking problems and higher omission rates is apparent (Table 4). Because areas with larger concentrations of minority populations are those with low mailback rates, and because these low rates are symptomatic of census-taking problems, we would strongly expect to find higher undercount rates in areas with low mailback rates.

As before, we have arranged district offices in order of decreasing mailback rates, and we see that as mailback rates decline, omission rates increase. The correlation between mailback rates and omission rates, $r = -.66$, is stronger than the correlation between mailback rates and rates of remaining unclassified ($r = -.48$). Problems were especially acute in district office 2 252 where the highest omission rates were observed. Turning to the lists, we see that the omission rate, like the rate of remaining unclassified, was highest for the confidential list, and lowest for the list of school children. However, the difference in rates between the billpayers and school children is negligible, and the omission rate for persons on the Medicaid Eligibility File is halfway between these two and that for the confidential list.

Finally, we see that omission rates in local areas have little to do with the composition of the areas in terms of the lists. The expected omission rates hardly varied, as they all fell in an interval between 8.0 and 8.4 per cent. The deviations between actual and expected omission rates tend to be positive where mail-

Table 4. Relationship Between Mailback and Omission Rates

District office	Mailback rate (%)	Omission rates in percent List[1] 1	2	3	4	Merged list	Expected rate[2]
2 246	85	7.8	8.0	9.2	6.2	8.2	8.1(+0.1)
2 203	82	6.0	6.5	4.1	6.7	5.4	8.1(−2.7)
2 201	79	6.3	11.0	2.6	4.6	5.1	8.1(−3.0)
2 202	79	8.0	9.4	3.9	3.1	6.0	8.1(−2.1)
2 242	78	6.6	5.0	3.8	0.0	5.1	8.1(−3.0)
2 255	78	6.1	6.5	6.9	3.4	6.2	8.2(−2.0)
2 243	77	10.0	4.0	9.7	0.0	8.3	8.2(+0.1)
2 240	76	11.4	8.4	12.6	14.3	10.9	8.2(+2.7)
2 251	73	9.8	6.8	5.3	7.4	7.8	8.3(−0.5)
2 253	72	15.9	6.9	11.3	10.3	10.9	8.1(+2.8)
2 245	70	17.0	6.0	12.2	3.4	10.4	8.3(+2.1)
2 247	70	8.2	6.3	4.7	7.8	6.6	8.1(−1.5)
2 256	70	10.2	12.7	6.8	11.2	9.9	8.2(+1.7)
2 244	68	10.5	8.5	9.2	6.5	9.6	8.4(+1.2)
2 250	68	10.7	13.9	7.1	8.4	10.6	8.4(+2.2)
2 248	67	15.0	4.9	10.9	6.5	8.9	8.2(+0.7)
2 249	67	5.4	5.6	6.5	4.3	5.7	8.1(−2.4)
2 241	64	13.1	7.1	5.1	7.8	9.5	8.4(+1.1)
2 254	63	16.7	7.8	11.3	6.9	12.0	8.0(+4.0)
2 252	60	18.9	16.7	16.7	21.9	17.9	8.1(+9.8)
Total city	74	9.6	8.0	6.7	6.6	8.2	

[1] Lists are defined as in Table 3.
[2] Expected rate calculated as in Table 3 with omission rates replacing rates of remaining undetermined.

back rates are low and negative where they are high, indicating the existence of local area effects beyond the composition of lists. Putting the results of Tables 3 and 4 together, we see that lists and areas with more hard-to-count people produced more problems of matching. Although the confidential list was harder to use, we found it to include a substantial number of omitted persons. Taking all the lists into account, we conclude that the list matching procedure had some success in identifying omitted persons in hard-to-count areas.

9. Evaluation of Procedure: Comparison With the Post Enumeration Program

One obvious problem with our procedure is that our combination of lists included approximately 5.7 million New Yorkers after persons found not to be living in the city on census day have been removed. Since there were 7.1 million persons counted, there are many more persons not on the lists to be accounted for. It would be convenient to be able to assume that omission rates of persons not on the list were the same as those rates for persons on the lists. This is a strong assumption which we would prefer not to make without supporting evidence.

We gain some insight into the reasonableness of the assumption by comparing the results of our megamatch procedure with those of the Post Enumeration Program in New York City. If the results are similar, we would have some reason to think that omission rates of persons not included on our lists were not greatly different from the rates of listed persons. In the PEP, the Census Bureau computed five separate omission rate estimates for New York City, and these ranged from 6.9 to 11.6 per cent. The main factor causing this variation was the Bureau's treatment of unresolved cases, which comprised about eight per cent of responding cases in the city PEP sample. For three of the estimates, the Census Bureau linked each unresolved case to a resolved case having similar personal characteristics, and imputed the status of counted or omitted of the resolved case to the unresolved case. These three omission rate estimates varied from 9.2 to 11.6 per cent. Just over half the unresolved cases were imputed to be omissions. At the other extreme, no imputations took place, and the unresolved cases were assumed to have been missed at the same rate as the rest of the sample. This had the effect of ignoring important geographic and other characteristics of the unresolved cases as well as the fact that no match had been found in the census.

On the megamatch study, we have seen that only about eight per cent of cases were unresolved. We also note that the megamatch omission rate of 8.1 per cent of the resolved cases is greater than the 6.9 per cent obtained when the unresolved cases were ignored on the PEP. If even half of the eight per cent of unresolved cases on the megamatch were omissions, the estimated omission rate would have been 12 per cent, which is comparable to the higher estimates provided by PEP. If the rate were 70 per cent, likely to be an upper bound, the omission rate estimate would be nearly 14 per cent. The omission rate of New Yorkers included in the compendium of lists is probably in the interval of 11 to 14 per cent.

We conclude that the megamatch estimates are comparable to those of the PEP. The megamatch method has one clear advantage, in that the more complete information provided concerning sample members seems to make matching easier. This is demonstrated by the fact that about eight per cent of cases on both studies were unresolved, even though matching on the megamatch took place two years later than matching on the PEP.

On a second criterion, the ability of lists to include hard to count people, the megamatch may be better. Census Bureau (1978, pp. 41,

62) studies indicate that in addition to a national noninterview rate of four per cent, about ten per cent of Blacks are not covered by the Current Population Survey. Given the well known problems of interviewing in New York City, it is likely that both rates are higher there. The megamatch lists are designed to include persons such as the unemployed, arraignees, and those eligible for Medicaid. Many young adult Black and Hispanic males and undocumented aliens are licensed drivers or electricity billpayers, and it seems reasonable to believe that many homeless people are eligible for Medicaid. It is likely that inclusion of these groups in the megalist increased the observed omission rate.

On a third criterion, inclusiveness, the PEP is better. The megamatch list included no more than 80 per cent of New Yorkers, and our study leaves us wondering who the 20 per cent of excluded persons might be. Consideration of the lists themselves provides some clues. Some of the omitted persons might not be particularly hard-to-count. Pre-school and parochial children not on welfare are two such groups and drivers who had obtained their licenses more than two years before the census are a third. On the other hand, unemployed persons who were no longer receiving benefits nor having driver's licenses or electric bills would be omitted. Thinking of all these groups at once, it is not obvious whether the people not included in the megalist are harder or easier to count than persons who were included in the list.

We can perhaps gain some insight by comparing the distribution, by district office, of New Yorkers counted in the census with those included in the megalist (Table 5). If the megalist overrepresents those areas with low mailback rates, we might think that it overestimated the omission rate, but we would draw the opposite conclusion if the megalist population was concentrated in areas with high mailback rates. In fact, megalist and counted populations are distributed similarly, with the megalist being slightly sparse in areas with high (above 80) or low (below 70) mailback rates. The differences are very small, and unlikely to have affected the omission rate. Indeed, if we weight the omission rate estimates found for the individual district offices (Table 4) by the proportions of the counted New York City population who live in the district, the resulting omission rate estimate declines only slightly, to 8.05 per cent. This would be the overall omission rate estimate if we assumed that the chances of being counted in the census are independent of being included in the megalist.

Table 5. Relationship Between Mailback Rates and Distributions of List and Counted New York City Populations

Range of mailback rates	Per cent of list population[1]	Per cent of counted city population	Aggregated omission rate[2]
80 and over	12.2	14.1	6.4
75 to 79	37.9	37.5	6.5
70 to 74	23.5	21.2	9.1
60 to 69	26.4	27.2	10.5
Total	100.0	100.0	8.1

[1] The list population has the cases remaining unclassified by the Census Bureau removed.
[2] Omission rates are calculated across all district offices in the category.

10. Conclusion

We developed the megamatch as an alternative to the PEP in a city where census-taking problems are great and omission rates hard to estimate. We were concerned that the CPS, given its well known undercoverage of Blacks, might not include sufficient numbers of the hard-to-count to provide an accurate estimate of the omission rate, and that matching problems might be substantial. The megamatch does seem to be a feasible method for estimating omission rates for a local area. Our problem is that the lists omitted over 20 per cent of New Yorkers, leaving us with an uncomfortably large amount of uncertainty. To establish whether the megamatch method is feasible on either a local or national basis, more research needs to be done, and we suggest three projects. The first is to look for national lists that together would include 90 to 100 per cent of the population. Some suggested lists for the United States are income tax payers and exemptions, food stamp recipients, and persons registering for the draft. Undocumented aliens may be identified from a list of persons sending money orders overseas.

The second is to compare the omission rate estimates provided by the megamatch with those provided by the Current Population Survey. The Census Bureau has data from both sources and could make such a comparison on a district office basis in New York City. The third is to study the comparative costs of the megamatch procedure with that of the Post Enumeration Program. We did not keep detailed cost information, but a skilled computer programmer did spend several months obtaining lists, sampling from them, and matching them against each other. This is a cost component not included in the PEP studies. On the other hand, we have heard no reports that the costs of field work and matching differ between the megamatch and PEP procedures. Finally, the cost of sampling is present in the PEP procedure unless an already existing sample such as the CPS is used. If the Census Bureau in 1990 plans to select an independent sample for estimating omission rates, then this will incur substantially greater costs than would our megamatch procedure.

In our view, the best strategy is to combine the megamatch and PEP-type survey procedures. We suggest using the Current Population Survey sample as the last list, thus incorporating it into the megamatch. Because the CPS sample is derived from lists of housing units, and not persons, it would not be possible to match subsequent lists against the CPS. However, persons in the CPS sample could be matched against the earlier lists. With the CPS in hand, we can limit the set of administrative lists to a small number, preferably less than five. If the lists focused on hard-to-count populations, we could realize the advantages of the inclusiveness of the CPS and the hard-to-count supplements without having to face the burdens of working with large numbers of lists. In New York City, we might restrict the supplements to the Medicaid Eligibility File and either licensed drivers or voters. With fewer lists to work with, there would be fewer matching problems and the project could be completed in a more timely fashion.

Our strategy would lessen a second, more political, problem as well. It took a lot of work to obtain the ten lists. Many New York City agencies were justifiably concerned about the confidentiality of their lists, and turned them over to us only after we agreed to merge them with other lists to reduce the likelihood of individuals being identified. With fewer lists, problems of obtaining them and matching samples would be reduced substantially, and greater resources could be allocated to field work. To summarize, we have not solved all the problems, but we conclude that the megamatch is potentially a good method for estimating omission rates. We hope that further work can be done with administrative records in the United States and elsewhere.

11. References

Alvey, W. and Scheuren, F. (1982): Background for an Administrative Records Census. Proceedings of the Social Statistics Section, American Statistical Association, pp. 137–146.

Bailar, B. (1983): "Affidavit," submitted to U.S. District Court, Southern District of New York. In *Cuomo v. Baldrige,* 80 Civ. 4550(JES).

Cowan, C. and Bettin, P.J. (1982): Estimates and Missing Data Problems in the Post Enumeration Program. Technical report, U.S. Bureau of the Census, October.

Cowan, C. and Fay, R.E. (1984): Estimates of Undercount in the 1980 Census. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 566–571.

Ericksen, E.P. (1983): "Affidavit," submitted to U.S. District Court, Southern District of New York. In *Cuomo v. Baldrige,* 80 Civ. 4550(JES).

Ericksen, E.P. and Kadane, J.B. (1985): Estimating the Population in a Census Year. Journal of the American Statistical Association, 80, pp. 98–131 (including Comments and Rejoinder).

Ericksen, E.P. and Kadane, J.B. (1987): Sensitivity Analysis of Estimates of Local Undercount in the 1980 U.S. Census. In R. Platek, J.N.K. Rao, C.E. Särndal, and M.P. Singh: Small Area Statistics: An International Symposium, pp. 22–45. John Wiley, New York.

Fellegi, I. (1980): Should the Census Count be Adjusted for Allocation Purposes: Equity Considerations. Conference on Census Undercount, U.S. Government Printing Office, Washington, D.C., pp. 193–203.

Fisher, F.M. (1984): "Testimony," given in U.S. District Court, Southern District of New York. In *Cuomo v. Baldrige,* 80 Civ. 4550(JES).

Freedman, D. and Navidi, W. (1986): Regression Models for Adjusting the 1980 Census. Statistical Science, 1, pp. 3–39 (including Comments and Rejoinder).

Kadane, J.B. and Lehoczky, J.P. (1976): Random Juror Selection from Multiple Lists. Operations Research, 24, pp. 207–219.

Karcher v. Daggett, 580 F.Supp. 1259, (1983).

Keyfitz, N. (1979): Information and Allocation: Two Uses of the 1980 Census. The American Statistician, 33, pp. 45–50.

National Research Council (1985): The Bicentennial Census: New Directions for Methodology in 1990. National Academy Press, Washington, D.C.

Redfern, P. (1983): A Study of the Future of Censuses of Population – Future Approaches. Unpublished paper commissioned by the Statistical Office of the European Communities.

Trussell, J. (1981): Should State and Local Area Census Counts be Adjusted? Population Index, 47, pp. 4–12.

Tukey, J.W. (1983): "Affidavit," submitted to U.S.District Court, Southern District of New York. In *Cuomo v. Baldrige,* 80 Civ. 4550(JES).

U.S.Bureau of the Census (1971): Testing Census Coverage Through Drivers' Licenses. 1970 Census Preliminary Evaluation Results Memorandum Series No. 21, Washington, D.C.

U.S.Bureau of the Census (1973): The Medicare Record Check: An Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census. In Census of Population and Housing: 1970 Research and Evaluation Program PHC(E)–5, U.S. Government Printing Office, Washington, D.C.

U.S.Bureau of the Census (1978): The Current Population Survey: Design and Methodology. Technical Paper 40, U.S. Government Printing Office, Washington, D.C.

U.S.Bureau of the Census (1982a): Coverage of the National Population in the 1980 Census by Age, Sex, and Race. Current Population Reports, Series P-23, No. 115, U.S. Government Printing Office, Washington, D.C.

U.S.Bureau of the Census (1982b): "Report by Census Bureau in Response to the Order of the Court Dated August 6, 1982," submitted November 12, 1982 to U.S.District Court, Southern District of New York. In *Cuomo v. Baldrige,* 80 Civ. 4550(JES).

U.S.Bureau of the Census (1985): The Coverage of Housing in the 1980 Census. In 1980 Census of Population and Housing Evaluation and Research Reports PHC80-E1, U.S. Government Printing Office, Washington, D.C.

Received February 1986
Revised November 1986