MATCHING IRS RECORDS TO CENSUS RECORDS: SOME PROBLEMS AND RESULTS

Danny R. Childers and Howard Hogan, Bureau of the Census

A. INTRODUCTION

This project has two principal aims: to investigate the feasibility of using the Internal Revenue Service Individual Master File (IMF) as a frame for matching to the census in order to estimate gross undercoverage in the census, and to study the difficulties in tracing individuals to the census using the IMF address.

There has been much discussion recently about using administrative records as a tool in evaluating the census. A combination of administrative records, such as IRS, Medicare, birth and death records, welfare, and other records, has been termed the megalist approach. For this study we selected a sample of persons who filed 1979 tax returns in April 1980. These joint and single filers were matched to the 1980 Decennial Census. Thus, this study can be considered to be a test of the IRS portion of the megalist for the working age population 18 to 65 years of age.

There are several possible advantages in using the IMF as the frame from which to draw a sample which will be independent of the census. Since it is not based on household interviews, it is unlikely to reproduce the same omissions as the census. It is especially good for groups with traditionally poor census coverage, such as young working age males. Sampling can be easily controlled on race and income thus permitting the over sampling of Black, Hispanic, poor, or other "hard to enumerate" groups.

Tracing is a key activity in the proposals for census coverage evaluation research. However, tracing is expensive and time consuming. Tracing will rely heavily on the use of administrative files, such as the IMF, which will be used to locate a more recent address. The IRS/Census match uses the IMF directly to obtain a census day residence address to match to the census. It thus increases the understanding of the IRS/IMF as an important tracing tool.

The match from IRS records to census records is conceptually simple. A sample of 10,887 primary and secondary filers was drawn from the 1979 IRS tax return file. The listing included the name and address of the taxpayer and spouse. The addresses were then coded to census geography and a search of census records was made to see if the sample persons were enumerated in the 1980 Decennial Census. If a person was not enumerated at the tax return address or if we could not code the address to census geography, the sample person was contacted to obtain a correct address or to determine if there existed another address at which the person was enumerated. The sample percentage unmatched will be used as an estimate of census omissions for the working age population.

The estimates of gross percent enumerated from this study cannot be compared to the net undercount estimates from Demographic Analysis or the Post Enumeration Program. This study was not designed to get another set of coverage estimates for the 1980 Decennial Census. It was a research effort to better understand tracing and matching techniques and to investigate the use of the IMF address as a starting point for matching to the census and tracing the initial not matched persons to their present address to obtain their 1980 census day residence.

B. Results

B.1 Estimated Percent Not Matched

The percent not matched in the age, race, and sex categories in Table 1 have been calculated as described in the noninterview adjustment section. The percent not matched in each category is an estimate of the gross missed rate. This number is high in each category partly because the followup interviewing was completed almost three and a half years after the census. This delay made it difficult to get accurate census day addresses, thus increasing the recall bias.

Table 1: Percent Not Matched for Race, Sex and Age

Race and Sex	Age 18-24	25-34	35-44	45-54	55-64	Total
Non Black Non Hispanic						
Male	25.6	20.1	6.9	4.2	3.5	13.4
Female	22.4	7.4	7.2	2.9	4.2	8.8
Total	24.1	14.0	7.0	3.5	3.8	11.1
Black Non Hispanic						
Male	33.9	36.6	20.3	15.2	11.4	26.6
Female	24.4	20.6	10.9	14.8	3.5	16.7
Total	29.3	28.1	15.3	15.0	8.1	21.5
Hispanic						
Male	17.6	19.0	13.4	18.6	21.2	17.6
Female	30.5	22.6	12.3	8.5	18.2	18.3
Total	24.5	20.9	12.8	12.3	20.0	19.3
Total						12.6

People younger than 35 had a higher percent not enumerated than those older than 35. After age 35 people seem to settle down more and a person who has a constant, fixed, and visible address has a better chance of being counted in the census. For Non Black, Non Hispanic and Black, Non Hispanic males the highest percent not enumerated was 18 to 34 and between 35 and 64 the percent not enumerated decreased as age increased. For Hispanic males the percent not enumerated seems to be relatively constant for all age groups. For females the percent not enumerated decreased as age increased in all race groups. Black, Non Hispanic and Hispanic females 18 to 34 had the highest percent not enumerated and Non Black, Non Hispanic females 18 to 24 had the highest percent not enumerated.

For persons who are Non Black, Non Hispanic and Black, Non Hispanic the percent not enumerated was

higher for males than for females. Hispanic females 18 to 34 have a higher percent not enumerated than males, but in the age groups 35 to 64 the males have a higher percent not enumerated than females.

The overall percent not enumerated was 12.6 for all races age 18 to 64. The percent not enumerated was 11.1 for Non Black, Non Hispanic; 21.5 for Black, Non Hispanic; and 19.3 for Hispanics.

There was a small group (less than 2 percent) in the study that had no age, sex, or race information. This group with no characteristics was also matched to the 1980 census and had 26.9 percent not matched. This group contained primarily immigrants and other adults who recently entered the labor force. These persons are not represented in Table 1. It is obviously more difficult to trace and match persons without characteristics.

The percent not enumerated in race, return type, and income categories has also been calculated in Table 2.

Table 2: Percent Not Matched for Race, Return Type and Income

Race and Type of Return	Income			
	Less than 8,000	8,000-14,999	15,000 or more	Total
Non Black Non Hispanic				
Joint	16.3	4.5	6.0	6.7
Single	24.1	20.8	7.8	20.8
Black Non Hispanic				
Joint	28.4	19.7	4.2	11.6
Single	29.8	30.3	36.2	30.9
Hispanic				
Joint	25.9	27.0	10.9	18.3
Single	25.5	24.6	33.6	26.3

The percent not matched were also calculated as described in the noninterview adjustment section. There are three income categories, less than $8,000, $8,000 to $14,999, and $15,000 and over, indicating gross income from the 1979 tax return. The income on the joint return is the total income for both filers and the income on the single return is the income for the single filer. The median household income is estimated by the Bureau from the 1980 census to be $16,841. Thus the $15,000 and over category is approximately median income and above.

The percent not matched for single filers in all three race categories is much higher than for joint filers. The joint filers are generally older and more settled than the single filers. A person who is single, young, and below the median income has a tendency to be more mobile than the remainder of the population. This person has a greater chance to be missed in the census. If that person is also Black or Hispanic, that person has an even greater chance of being missed in the census. Joint filers with a more permanent, fixed, and visible address will be missed at a lower rate.

For joint filers the percent not enumerated had a downward trend as income increased. Hispanic joint filers had the highest percent not enumerated, Black, Non Hispanic was next, and Non Black, Non Hispanic had the lowest percent not enumerated. For the Non Black, Non Hispanic single filers the percent not enumerated decreased as income increased. For Black, Non Hispanic and Hispanic single filers the percent not enumerated increased as income increased.

B.2 Percent Not Traced

One objective of this matching study using the IMF as a sampling frame was to see what proportion of the sample persons could not be traced to their place of residence for a followup interview. If a high proportion of the persons selected from the 1979 tax return file could not be matched at the IMF address and could not be traced to their present address during all of the three followup attempts, the IRS/IMF would not be a good source for sampling persons for coverage evaluation of the census. On the other hand, the 1979 tax return was filed before April 15, 1980 and should be a good source for sampling the working age population that is 18 to 64 years of age.

Table 3: Percent Not Traced

Race and Sex	Age					
	18-24	25-34	35-44	45-54	55-64	Total
Non Black Non Hispanic						
Male	4.4	6.1	2.9	1.7	0.0	3.4
Female	5.1	3.3	1.2	0.0	0.0	2.1
Total	4.7	4.7	2.0	0.8	0.0	2.8
Black Non Hispanic						
Male	4.0	6.7	10.1	0.0	0.0	5.2
Female	0.0	3.9	14.5	0.0	0.0	4.9
Total	2.0	5.2	12.5	0.0	0.0	5.0
Hispanic						
Male	3.9	8.7	0.0	8.7	0.0	5.7
Female	0.0	16.0	3.0	4.3	0.0	6.5
Total	1.8	12.5	1.5	6.5	0.0	5.8
Total						
Male	4.3	6.3	3.4	1.9	0.0	3.7
Female	4.3	4.1	2.6	0.2	0.0	2.6
Total	4.3	5.2	3.0	1.0	0.0	3.1

A sample person was coded as "tracing failed" after the mail followup questionnaire was not returned, the telephone interviewer could not locate a telephone number for the sample person or anyone who ever heard of him/her, and the field interviewer was unable to find the sample person or anyone who could give any information about the sample person.

The estimated percent not traced for the total working age population was 3.1 percent (see Table 3). The percent not traced is 2.8 for Non Black, Non Hispanic, 5.0 for Black, Non Hispanic, and 5.8 for Hispanic. For Non Black, Non Hispanic and for Black, Non Hispanic the percent not traced was higher for males than for females. For Hispanics the percent not traced was higher for females than for males.

C. Tracing and Matching

C.1 Initial Matching at 1979 IRS/IMF Address

An initial screening removed the IRS cases with addresses that were post office boxes, rural routes, military addresses, and other nonstandard addresses that could not be easily geocoded. An attempt was made to code the remaining cases to 1980 census geography. The cases that could not be coded to census geography were removed and combined with the ones removed during the initial screening resulting in 1751 cases that were not assigned census geography. The remaining 5685 cases with census geography were matched to the 1980 census. Of the 5685 cases searched at the IRS/IMF address the single filer or both filers on the joint return were matched 78.2 percent of the time to the 1980 Decennial Census at the address reported on the 1979 IRS return. An additional 82 cases had one of the filers on a joint return matched at the IRS address, (i.e., partially matched). Also, 41 cases were determined to be ineligible to be included in the census, because they were deceased or living out of the country. They were coded as such and removed from further processing. Thus, there were 4485 returns completely coded after this phase of procesing or 60.3 percent of the IRS sample of single and joint returns.

The 4485 joint and single returns that were coded during the initial match contained 6826 sample persons. Both filers on a joint return were in the sample. Almost all of the resolved cases (99 percent) were matched, because only a few were allowed to be anything else before followup.

C.2 Prefollowup Sorting

After the initial match, all cases were assigned to one of three groups:

A. Cases that could be assigned a final match status without followup.
B. Cases where the address could not be located in the census.
C. Cases where the address was found in the census, but the sample people were not found.

Group A needed no followup because the final match status could be assigned after matching the sample persons at the 1980 tax return address to the census questionnaire for that address. Group B went to followup because the housing unit reported to IRS either could not be converted to census geography or no attempt was made to geocode the address because it was rural or vague. These cases were sent to followup, first, to determine the census day address and second, to get a location description that would enable us to convert the address to census geography. Group C was geocoded and matched to the census questionnaire for the address reported on the 1980 IRS tax return, but the sample persons were not listed on the 1980 census questionnaire. These cases were sent to followup to get the 1980 census day address.

The reasoning behind creating the above three groups was that the nonmatch rate would be different in the three groups. The noninterview adjustment was conducted separately within the three groups in each demographic subgroup.

The number of sample persons in each prefollowup category and in each region are in Table 4. The numbers in parentheses are the percent of persons in each region or total assigned each prefollowup code. Thus, 62.5 percent of the sample persons were assigned a final match status without followup. Also a total of 21.0 percent of the sample persons were followed up because the address could not be located in the census or was rural or vague and 16.6 percent were followed up because the address was found in the census, but without the sample persons listed on the census questionnaire for the address.

The South had a higher percent of sample persons with prefollowup code B, because the South had more rural addresses and more addresses that were hard to assign census geography. The West contained slightly more sample persons where the address was matched, but the sample persons were not (i.e. code C), indicating that there were more movers in the West than in the other three regions.

Table 4: Prefollowup Code by Region

Pre-followup Code	Region NE	S	NC	W	Total
A	1542 (.667)	1959 (.543)	1543 (.662)	1753 (.666)	6797 (.625)
B	371 (.161)	1111 (.308)	424 (.182)	379 (.144)	2285 (.210)
C	398 (.172)	543 (.150)	365 (.157)	499 (.190)	1805 (.166)
Total	2311	3613	2332	2631	10887

The type B addresses were defined as follows:

B1: An address that was rural or vague (i.e. Post Office boxes or rural routes).
B2: A complete address that could not be coded to census geography without more location information, such as cross streets, neighbor's names, sketch map, and a location description.
B3: An address that was coded to census geography, but after searching for the address in the census, the address appeared to be incorrectly coded. It was sent to followup for more information.
B4: An address that appeared to be a business address, because the census area contained no living quarters.

B5: The address was correctly coded to census geography, but the address was not listed in the census. This may also be a business address.

B6: A military address that was not easily coded to census geography.

Table 5 contains the percent of type B addresses that are in the above six groups. Almost 80 percent of the addresses with prefollowup code B were rural or vague. The persons who use their business and their lawyer or accountant's address to file their taxes had been believed to be quite large. The categories for business addresses were B4 and B5. Together they accounted for 5.7 percent of the type B addresses and only 1.1 percent of the total addresses.

Table 5: Type B Addresses

Address Type	Percent of B Addresses	Percent of Total Addresses
B1	79.8	16.7
B2	9.9	2.1
B3	4.4	.9
B4	1.2	.2
B5	4.5	.9
B6	.3	.1
Total	100.0	21.0

C.3 Mail Followup

The 2951 unresolved cases after the initial match (39.7 percent of the sample returns) were sent mail followup questionnaires to obtain the sample person's address of residence on April 1, 1980. The 2951 cases involved in mail followup included the 1200 cases where one or more filers were unmatched after matching to the IRS address on the 1979 tax return and the 1751 cases that could not be easily coded to census geography. The cases initially classified as unable to code to census geography were sent a questionnaire designed to obtain the exact 1980 address before additional money, time, and effort were used to code these addresses to census geography. Also, for the post office boxes and rural addresses, a location description and neighboring addresses were requested to make the location of the 1980 residence on a map easier or possible in some cases. There was also a question on the form asking if two names were for the same person in cases where the filer's name was similar to the name listed on the census questionnaire.

The number of postmaster returns (14.2 percent) was expected, since many people have moved in the two years and six months since census day. The nonresponse rate (52.7 percent) was higher than anticipated and was disappointing. In only 24.8 percent of the cases was a useful reply received.

The persons followed up by mail separated by type of address before followup are in Table 6. Address type B indicates that the address could not be located in the census and type C indicates that the address was found, but the sample people were not found. More than twice as many persons with address type B returned a completed mail followup questionnaire than type C, because they did live at the tax return address. The type B addresses were difficult to convert to census geography. The post master return (PMR) rate was 16.7 percent for type C and 11.6 percent for type B. There was a higher PMR rate for type C because more of them had moved since 1980 than for the type B addresses.

Table 6: Persons in Mail Followup by Type of Address

Type of Reply	IRS/IMF Address Not Found Census (B)		IRS/IMF Address Found in Census but Sample Filers Not Listed (C)	
	Persons	% of Total	Persons	% of Total
Mail Reply	756	33.1	269	15.2
Post Master return	266	11.6	296	16.7
No return	1,182	51.7	953	53.7
Not recorded	81	3.5	255	14.4
Total	2,285		1,773	

A mail followup questionnaire was returned for 1005 persons and 936 persons were assigned a final enumeration status (see Table 7).

Table 7: Final Match Status for Persons Who Returned the Mail Followup Questionnaire

Match Status	Persons	Percent of Coded Persons
Coded	936	
Matched	729	77.9
Not Matched	147	15.7
Final Nonresponse	40	4.3
Out of Scope	20	2.1
Not Coded	69	
Total	1005	

C.4 Telephone Followup

All cases that returned a mail followup questionnaire, but required additional information, were sent to telephone followup. Many of these cases were ones where one filer was matched, but the spouse was not, because of divorce or separation. A response of "divorced" does not help to locate the census questionnaire. The exact census day address is needed. Others needed additional information because the mail followup form was not complete. Mailing addresses, either post office boxes or other rural addresses given by the respondent, were easier to geocode with additional location description and intersecting streets. Many college students or other younger persons without an address they consider as their permanent address will respond that their address on April 1, 1980 is their parents' address even when they did not live there. If a single filer was not listed

on a census questionnaire that was obviously their family's, the case was sent to telephone followup. A telephone interviewer is more able to discern the person's true census day address than the respondent on a mail followup form.

The postmaster returns and nonresponse cases were subsampled in order to reduce cost. One fourth of the cases where the characteristics were not available for the primary filer were sent to telephone followup along with one half of the remaining PMR and nonresponse cases.

The telephone followup had an overall success rate of 39.1 percent. The cases that were nonresponses during mail followup had a higher completion rate than the ones that were post master returns. Many of the sample persons who did not respond to the mail followup questionnaire, but still lived at the IRS filing address, would give the necessary information to the telephone interviewer.

A sample person was not traced if no one could be located by telephone to give us any information about the sample person and no telephone number could be obtained. If the telephone interviewer talked to the sample person or to someone who knew the sample person, a person was considered traced even if the information was not geocodeable or the person was classified an unresolved. In these instances, no useful information was obtained for locating the sample person, but going to the field would probably not obtain anything more useful. For example, if the sample person said that he moved around a lot in 1980 and did not remember where he was living on April 1, 1980, he was coded as unresolved after telephone followup. No field followup was done for these unresolved cases, since it is not likely that talking to him in person will yield any better information than conducting the interview over the telephone. Thus only untraced cases were eligible for field followup.

All sample persons who were traced during telephone followup were assigned a match status. The results of the match to the census are in Table 8.

Table 8: Final Match Status for Persons who were Traced During Telephone Followup

Match Status	Persons	% of Persons Traced
Traced	594	
Matched	286	48.1
Not Matched	194	32.7
Final Non Response	107	18.0
Out of Scope	7	1.2
Not Traced	925	
Total	1519	

C.5 Field Followup

One fourth of the untraced persons after telephone followup were sent to the field for a personal interview. All persons who were involved in field followup were searched in the census and final match codes were assigned (see Table 9). There were 187 persons traced and 53 persons not traced of the 240 persons in field followup.

Table 9: Final Match Status for Persons in Field Followup

Match Status	Persons	% of Persons Followed Up In the Field
Total	240	
Matched	75	31.2
Not Matched	86	35.8
Final Non Response	23	9.6
Out of Scope	3	1.2
Tracing failed	53	22.1

At each phase of the tracing and matching operations the final enumeration status was resolved for some sample persons. For others, the enumeration status could not be determined without additional followup. The percent matched and not matched of the resolved cases during the initial match of the sample persons at the 1979 IMF address and during each of the followup operations has been calculated in Table 10. As expected, the percent matched decreased with each additional operation. The percent matched is not constant because only the unresolved and untraced cases went to followup and as the followup progressed from mail to telephone to personal visit, the cases become increasingly more difficult and a higher percentage is truly not matched to the census.

Table 10: Percent Matched and Not Matched of Traced or Resolved Cases at Each Phase

Match Status	Initial Match	Mail Follow Up	Phone Follow Up	Field Follow Up
Percent Matched	98.9	78.0	48.1	40.1
Percent Not Matched	0.06	15.7	32.7	46.0

D. **Nonresponse Adjustment**

Nonresponse adjustment is normally based upon variables such as age, race, sex, and size of place. This study was designed to use the status of the housing unit after the initial match as another variable for nonresponse adjustment.

The nonresponse adjustment was done separately for each stage of followup within each cell group. In this study we tried to separate the cases into homogeneous groups for whom the percent not matched that were interviewed would be used as the estimate of the percent not matched for the noninterview cases. This resulted in an estimate of the percent not matched that is larger, but is believed to be closer to the actual percent not matched.

For estimates of percent not matched, the noninterview and the tracing failed cases were combined into one nonresponse category and the nonresponses were allocated to matched or not matched with each cell group for each stage of followup. For estimates of tracing failed the noninterviews were considered to be traced. A person

was considered traced even if the information that was received during followup was not useful in locating a questionnaire in the census. A person was classified as tracing failed after the mail followup questionnaire was not returned, the telephone interviewer could not locate a telephone number for the sample person or anyone who ever heard of him/her, and the field interviewer was unable to find the sample person or anyone who could give any information about the sample person.

If the nonresponse adjustment is done without the prefollowup and followup code classifications, the resulting percent not matched in each age, race, and sex category is in Table 11. The percent not matched is lower when ignoring the prefollowup and followup codes, but may not be as accurate.

Table 11: Percent Not Matched (ignoring Prefollowup and Followup codes)

Race and Sex	Age 18-24	25-34	35-44	45-54	55-64	Total
Non Black, Non Hispanic						
Male	21.0	14.4	5.3	4.0	3.3	9.8
Female	16.8	6.1	6.2	2.7	3.7	6.9
Total						8.4
Black, Non Hispanic						
Male	30.1	32.1	11.5	15.0	10.3	22.3
Female	24.7	16.1	2.6	13.9	3.6	13.3
Total						17.6
Hispanic						
Male	16.0	16.1	13.4	11.3	21.2	15.2
Female	30.0	16.0	10.5	9.5	18.2	16.7
Total						16.0

E. Comparison with the 1980 Post Enumeration Program

This research project was designed to study the IRS records as a source for sampling persons in the working age population. These persons were compared to the 1980 census using the address on the tax return that is filed in April 1980. If the joint and single filers were not found in the census at the tax return address, they were traced to their present address to obtain the address on April 1, 1980. Estimates of the gross percent missed were made for this study.

This study was not meant to be an alternate source of estimates of miss rates for evaluating the 1980 Decennial Census. For comparison purposes, the estimates of gross percent missed from the Post Enumeration Program are in Table 12. These figures are also subject to biases and other errors which in some cases may be quite large. They are given here merely to indicate general size.

Table 12: PEP Estimates of Gross Percent Missed

Race	Percent
White	5
Black	12
Not Spanish	5
Spanish	10

The estimates of gross percent missed in this study were higher, but the mail followup was done two and one half years after the census, the telephone followup was conducted in the spring of 1983, and field followup was attempted in August 1983. The recall bias would have been less if the study was done closer to census day. Also, since this research project was only for the working age population (i.e. 18-64), young persons and older persons were not included in the estimates of gross percent missed in this study.

The Bureau intends to continue research into the use of nonhousehold sources for coverage evaluation. This study has demonstrated that the problems of post office boxes, rural routes, and business addresses can be overcome with proper followup procedures. The ease of taking a large and diverse sample including many non traditional addresses was impressive. We believe that the potential of this sampling frame is immense.