Evaluation of 1996 Community Census Administrative Records File

Jay Kim, Elizabeth T. Huang and Kent Marquis
Jay Kim, Bureau of the Census, Suitland, MD 20233

Key Words: Administrative Records Database, Source File, Matching, Mean Absolute Relative Error

I. Introduction

Population and housing censuses are part of the nation's information infrastructure. But conventional census collection processes are expensive and burdensome to citizens and are becoming increasingly difficult to implement. Technology advances encourage examining whether administrative records, already part of the federal government information system, could be used either to improve or substitute for the conventional processes.

This report evaluates a specially-built administrative records database for Chicago by comparing information in the database to the 1996 Census test in Chicago. We compare counts and characteristics for households, people, addresses, blocks and entire test site. High match rates or agreement rates are desirable (see Buser, et al's (1998) for other test sites results).

Results consist of tables and short, accompanying discussions. The results and discussion illustrate that administrative records procedures can provide information needed for a population census but many issues must be addressed and solved before the information is considered accurate and complete.

The report is organized as follows. In section II, we describe the methodology for creating and evaluating the 1996 Administrative Records Database (ARD). In section III, we give the evaluation results. Summary and discussion are given in section IV.

II. Methodology

II.1. 1996 Administrative Records Database (ARD)

We outline the steps used to process the 1996 administrative records files and to create the ARD for Chicago test site.

Step 1 Acquisition of Files

The files we used include:

- Internal Revenue Service Tax Year 1994 Individual Master Tax Return File (IRS) with up to 4 dependents per tax filer unit
- Multifamily Tenant Characteristics System (MTCS), Department of Housing and Urban Development
- Tenant Rental Assistance Certification System (TRACS), Department of Housing and Urban Development
- Registration File, Selective Service System
- 1996 Medicaid file - Client Information System, Illinois Department of Public Aid
- 1996 Medicare Enrollment Database from Health Care Financing Administrative (HCFA)
- The Social Security Administration NUMIDENT file was used to validate the social security number and to gain missing demographic characteristics for people found on other files.

Step 2 Standardization, Geocoding, and Social Security Number Verification

Standardization of the content of data fields to be used in linking records is necessary to assure that correct matches are made. Standardization is a table look-up and substitution process that translates variations on the expression of an underlying value to a single (standard) version of the value. An example of name standardizing would be to substitute John for input values of Juan, Johann, Jean, Giouvani, John, Jon, etc. Address standardizing involved parsing the address string into a large number of components and assigning standard values for each non-null component, e.g., for the street direction prefix component, substitute S for South, So., etc.

This step was done separately for each file. It included the following 4 substeps:

1) Name standardization- names are standardized, a score is put on the output record for completeness of the name, and a code identifying which file the record is from is added.

2) Address standardization-addresses are standardized using the most recent version of the Geography Division address standardizer.

3) Geocoding - census geography is appended to each output record, including the TIGER (Topologically Integrated Geographic Encoding and Referencing System) Line ID. Also a Housing Unit Identification Number(HUID) is created for each record for all types of addresses.

4) Social Security Number (SSN) verification - each input record is matched with the NUMIDENT file to verify the validity of the SSN.

Step 3. Concatenation - All input administrative record files are concatenated to form one person records file. Records are dropped if the address is located outside of the site boundary or if the record lacked a person's name.

Step 4. Within HUID Unduplication - Based on a predefined order of selection, records which have high probabilities of matches on variables such as name, date of birth, gender, SSN, race, and Hispanic origin within the same HUID are merged into one person record using Winkler's matching algorithm.

Step 5. Address Independent Unduplication - Resulting person records with the same SSNs are identified across

HUID groups and are retained based on a priority scheme.

Step 6. Person records missing an address and those for people identified as deceased were dropped.

Step 7. All person records bearing the same HUID were combined to create ARD households.

II.2. Evaluation Methods

For evaluation of the ARD it was matched to the 1996 Community Census test data from the Chicago test site.

Two census extract files were used, one from the Decennial Master Address File (DMAF) and the other from the 1996 Community Census Unedited File (CUF). They are an address file and a person file with address and demographic data, respectively.

For the person match, the ARD was matched to the Census for the same HUID by name, date of birth and sex.

For the address match, the 1996 ARD was matched to the 1996 DMAF, using HUID as a match key. The full address match is the address match where street name, house number and apartment number are the same in both the census and ARD records. The basic street address match is the address match by house number, street name, but not the apartment unit number. HUID has 6 different address types: geocoded street addresses, P.O. Box addresses, geocoded property addresses, non-geocodable addresses, non-standardized addresses and no addresses on file. Only geocoded street addresses can be used for the address match.

For the whole household match, all ARD records for a household (defined by persons with the same HUID) must match to the Census household records by HUID and by name, date of birth and sex.

To measure the accuracy of the ARD data for the demographic characteristics of the matched persons or the unmatched households with same HUID, we used an agreement rate, the ratio of the number of agreement to the total. Two totals were used, and hence two ratios were calculated. One total includes all cases (including the missing values) and the other excluding the missing value. The first agreement rate is an absolute (unconditional) probability of agreement and the other agreement rate is conditional on the existing values on both ARD and the census.

To measure the accuracy of the aggregate population count from the ARD with the census count at the tract (block) level, we used two statistics:

(1) Number of tracts (blocks) with a positive difference

The number (or proportion) of tracts (blocks) where the difference of the ARD count and the census count at the tract (block) level is greater than zero.

(2) MARE - Mean Absolute Relative Error

MARE is the average of the absolute difference of the ARD count at tract (block) level with the census count relative to the average of the census count of tracts (blocks) in the test site.

Let a_i be the person count (or person count of a specified person demographic characteristic) of the i^{th} area (tract (block)) from the ARD record, i = 1,...,n. (There are 7 tracts).

Let c_i be the person count (or person count of a specified person demographic characteristic) of the i^{th} area (tract (block)) from census record, i = 1,..,n.

The error measurement of MARE is defined by

$$\text{MARE} = \{(1/n) \sum_{i=1}^{n} | a_i - c_i | \} / \{(1/n) \sum_{i=1}^{n} c_i\}$$

When we have the missing data say ARD tract (block) ID missing, or certain demographic characteristics missing in census or ARD person record, the MARE (after adjusting for missing data) can be calculated using the available n_r records by the following formula:

MARE (after adjusting for missing data)

$$= \{(1/n_r) \sum_{i=1}^{n_r} | a_i r_a - c_i r_c | \} / \{(1/n_r) \sum_{i=1}^{n_r} c_i r_c\}$$

where

r_a = Total ARD persons / (Total ARD persons - number of ARD persons missing)

r_c = Total census persons / (Total census persons - number of census persons missing)

The estimate of ARD (census) adjustment factor r_a (r_c) is varied by the demographic characteristic, because different characteristics have different missing rate.

III. Evaluation Results

In this section, we compare ARD information to census, (information for addresses, households, people, and characteristics of people) assuming that census information is correct.

Table 1 shows there are 8,654 addresses in the ARD, which is lower than the corresponding census address count (9,833). The ARD household count is the same as the address count, because only if resident(s) was present at an address, a household was formed. However, because of vacant housing units, the census household count is 2,338 less than the census address count. Note now the ARD household count outnumbers the census household count by 1,159. The ARD person count also is greater than the census person count by 4,908. The last row shows the match count for each of the categories observed above.

Table 1 ARD and Census Counts

	Address	Household	Persons
ARD	8,654	8,654	25,403
Census	9,833	7,495	20,495
Match	3,617	645	3,015

III. 1. Evaluation of Address Match

The ARD exact address match rate is 41.8 percent, i.e., out of 8,654 total ARD addresses, 3,617 matched to the census address list down to the subunit id. When matching criterion was relaxed by ignoring the subunit id, additional 29.1 percent (2,516/8,654) of the ARD addresses were matched. Note 27.5 percent (2,379/8,654) of the ARD addresses were P.O. Box type. Since the Bureau of the Census uses city style address only in urban areas such as Chicago, no P.O. Box address in the ARD could be matched to the census address.

We also looked at how well the addresses from different types of file sources matched to the census address database. IRS has the highest match rate with 48.2 percent. The second highest match rate was achieved by the Selective Service source files, with a percentage of 45.4, followed by Medicare and Medicaid with 35.1 and 26.9 percent, respectively. IRS contributed the highest number of matched addresses of any of the sources (3,222) but also the highest number of unmatched addresses (3,465).

III. 2 Evaluation of Whole Household Match

One of the goals of this evaluation is to measure improvements in our ability to create households using administrative records in the decennial census and for other applications. At the addresses which match between the ARD and the census, residents were compared between the files. When all persons match at the address, we call it a whole household (WHH) match. In this section, we examine the number (and percentage) of the ARD WHH matches and the characteristics of these matched households.

Table 1 shows 645 WHH matches, which amount to 7.5 percent (645/8,654) of the ARD households and 8.6 percent (645/7,495) of the census households.

Table 2 shows the number and rate ARD source files contributed to the 645 WHH matches. TRACS and IRS show the highest WHH match rates. The IRS file has the highest frequency of all the source files, followed by Medicare. Since some ARD households may have been created from more than one source file, the column total of columns 2 and 4 in Table 2 is greater than or equal to total WHH match and WHH count, respectively.

TABLE 2 - ARD WHH Matches by Source File

ARD Source File	WHH Match		Total
	Number	Percent	
Any IRS	586	8.76	6,687
Any Medicaid	29	1.54	1,889
Any Medicare	123	5.57	2,209
Any TRACS	10	27.78	38
Any MTCS	0	0.00	69
Any Selective Service	7	0.84	837

The WHH match rate for different ARD household sizes shows an interesting pattern. As the number of persons per household increases, the WHH match rate clearly decreases. The match rate is 11.5 percent (the highest) for one person ARD household and 0.1 percent (the lowest) for ARD households of 7 or more.

All White households have the highest WHH match rate with 14.9 percent. All Black households have 2.6 percent match rate and all Other have 2 percent match rate.

A total of 11.8 percent of all ARD nonHispanic households are WHH matches, whereas all Hispanic and mixed were less than one and a half percent of WHH matches.

The highest WHH match rates are found in households with 1 or 2 adults and no children. Households with 1 or 2 adults in the 31-64 age group in this category show a WHH match rate of 18.2 percent, followed by a 12.5 percent households with 1 or 2 adults aged 65 and older. The poorest match rates were obtained for households with the '3 or more adults with children'.

III.3. Comparison of Household Demographic Characteristics for Non-Whole Household Matches between ARD and the Census

Non-whole household matches refer to the address-defined groups whose addresses match the census addresses but their members do not match the census household members completely. That is, either some of the members match or none of the members match to the census. This analysis can shed some light on how successfully the household substitution can be done if the Bureau decides to use administrative records households information for imputing the characteristic of nonresponding households (cold-deck approach), rather than the current hot-deck approach.

There are 2,592 non-whole household matches. Comparing the household sizes between the ARD and the Census, we were able to correctly classify 1,276 households. This amounts to 49.2 percent of total. Thus even if the household members do not match completely between the files, around 50 percent of the time ARD gives the correct household size.

If all household members in the same household are White, Black or Indian, the household was classified as "all White," "all Black" or "all Indian," respectively. If race value were missing for all household members of a household, the household was classified as "missing." The rest was classified as "Mixed, other." Thus the last category includes, for example, the households whose members are "all Asian," "all Others," mixture of different races or mixture of some race and missing value for race. The number of households whose race composition matches between the files is 1,024. It is 39.5 percent of all households considered. This is the absolute race agreement rate. If "all Missing" and "mixed" are excluded from total, the percentage goes up to 95.6 percent. Similarly, the absolute and conditional Hispanic origin agreement rate are 33.5 and 88.8 percent, respectively.

III.4. Evaluation of Matched Person Records

In this section, we compare demographic characteristics for the matched persons between the files. This is to check the

accuracy of the demographic data for the matched ARD persons. Since name, date of birth and gender were used for matching persons between ARD and the census, they should be of good quality for the persons. Thus they will be ignored and only race and Hispanic origin will be considered here. We will also investigate in this section which administrative file(s) or combinations of files contribute to the matched person.

III.4.1 Evaluation of Matched Person's Demographic Characteristics

Absolute and conditional race agreement rates are 55.3 and 88.1 percent, respectively. The rates are 71.3 and 88.7 percent, respectively, for Hispanic origin.

III..4.2 Evaluation of Demographic Data for Matched Persons by Source Files

Among the final ARD person records, information may originate from a single source or from multiple sources. Most final person records (91.9 percent) are from a single source. The majority of the single source, ARD person records (61.8 percent) are from the IRS file. Medicaid and Medicare together account for 28.0 percent of all final ARD person records. It is interesting to note that neither MTCS nor TRACS accounts for a single ARD person record (single source). Note that 8.0 percent of the records are from two source files and .2 percent from three sources.

Among 3,015 matched persons, a total of 86.6 percent of all matches are from single source files, 13.1 percent is from two source files and the rest (.3 percent) from three sources. IRS as a single source accounts for 70.9 percent of all matched persons. The share of Medicaid and Medicare is 9.8 and 5.3 percent, respectively.

Overall single source files has a match rate of 11.2 percent. The match rates for all combinations of two source files and three source files are 19.5 and 22.7 percent, respectively. Thus it can be observed that when the record is contributed by more files, the record's chance of matching goes up.

Among the single source files, the IRS file has the highest match rate (13.6 percent) and the Medicaid file has the next highest rate (7.7 percent), followed by Selective Service (5.2 percent) and Medicare (4.9 percent).

III.5. Comparison of the ARD Total with the Census Total at the Tract and Block Level.

A total of 4,869 ARD person records were missing tract (block) IDs. For persons with nonmissing tract (block) IDs, the ARD has more people than the census in 4 of 7 (57 percent) tracts. In addition, the ARD has more people than census in seven tracts for the following demographic characteristics: males (57 percent), females (71 percent), blacks (85 percent), not Hispanic (86 percent), and the 2 extreme age groups (less than 19 years (86 percent) and 65 or over (100 percent)).

MARE at tract level is 0.16 without missing data adjustment. In other words, the average tract difference (ignoring sign) between the ARD and the census counts is about 16 percent of the average of census count for tracts. If we adjusted for the missing data, the MARE for tract is 0.26.

The age group of 31 to 64 years old has the smallest MARE whether the missing data are adjusted or not.

The highest MARE, 1.24, at the tract level is for black persons. In all but one tract, the ARD was able to count more black people than the census. The missing race for ARD is 38.8 percent of the ARD person records with tract IDs; for census is 19.4 percent of census person records. The MARE for black persons after missing data adjustment is 2.63.

For Hispanic origin, the MARE at the tract level is 0.29, and the MARE after adjusting for the missing data is 0.17. No tract in all 7 tracts has the total ARD Hispanic persons greater than the total of census Hispanic persons. It appears that ARD undercounts the Hispanic persons in the tract. The missing Hispanic origin rate is 14 percent of the ARD person records with tract IDs; and 18.7 percent of the census person records.

The MARE for the age group of less than 19 years old is 0.15 without missing data adjustment; and 0.36 with the missing data adjustment. The low MARE for age less than 19 years suggests that the inclusion of the dependents from the IRS file enhances the quality of the ARD coverage of the age group less than 19. However in the WHH matched households, only 6.67 percent are households with children. It seems that the coverage of total children in ARD is not so much a problem as the "right" children at the "right" address.

The second highest MARE (1.13) at the tract level is for the age group of over 65. (MARE after adjusting for missing data is 1.56). Ignoring the missing data, all 7 tracts have more ARD persons of age over 65 than the census. The over coverage of the old age persons (65+) in ARD may be due to using the Medicare file as a source of ARD (a total of 12.90 percent of the ARD persons is from the Medicare file).

Without adjusting the missing data the smallest MARE (0.08) at tract level is for people in the age group of 31-64. Only in one tract is the ARD total for this age group greater than the corresponding census total. The MARE after adjusting for missing data is 0.13.

In general, and as expected the block level MARE is larger than the tract level for the same demographic characteristic. For the population count, the block level MARE is 0.26 (the MARE after adjusting missing data is 0.40), and the number (or the proportion) of blocks with the ARD greater than the census count is 77 out of 106 (72 percent).

At the block level, the demographic characteristics for which the unadjusted ARD person counts exceed the census person counts are: all persons (72 percent of blocks), male (73 percent of blocks), female (66 percent of blocks), black (67 percent of blocks), not Hispanic (94 percent of blocks), age group <19 (62 percent of blocks), age group 65+ (93 percent of blocks).

The ARD had fewer block counts greater than the census for these demographic characteristics: white (38 percent of blocks), other race (includes American Indian and Asian Pacific) (10 percent of blocks), Hispanic (7 percent of blocks).

The largest MARE at the block level is 1.41 for black persons or 2.67 with adjusted missing data. A total of 1,384 black persons in the census and 3,089 in the ARD. The average block total of black persons is 13 for census, and 29 for ARD. The ARD has substantially more blacks than census with or without adjusting for missing block information.

For Hispanic origin, there are 6,950 ARD Hispanic persons and 9,810 Census Hispanic persons. The average Hispanic

persons in a block is 65 for ARD and 92 for census. The MARE for Hispanic at the block level is 0.30, and the MARE after adjusting the missing data is 0.20. In only 6 percent of the blocks, the ARD total Hispanic person is greater than the census at the block level.

The second largest MARE at block level is 1.32 for the age group of 65 above or 1.74 for adjusted data . The average block count for the age group of 65 above is 17 for census and 34 for ARD. There are 99 blocks out of 106 blocks (93 percent) having ARD total greater than the census total for the age group of 65 above. There are substantially more old age (65+) persons in ARD than census.

The smallest MARE at the block level is 0.17 for the age group of 31 to 64 without adjusting for missing data or 0.21 for adjusted data. There are 45 blocks from a total of 106 blocks (42 percent) where the ARD block totals of age 31 to 64 are greater than the corresponding age group total for census block. The second smallest MARE at the block level is 0.24 for the age group of less than 19 years old before missing data adjustment. In 62 percent of the blocks the ARD has more children (< 19) than the census.

IV. Discussion and Recommendations for Research

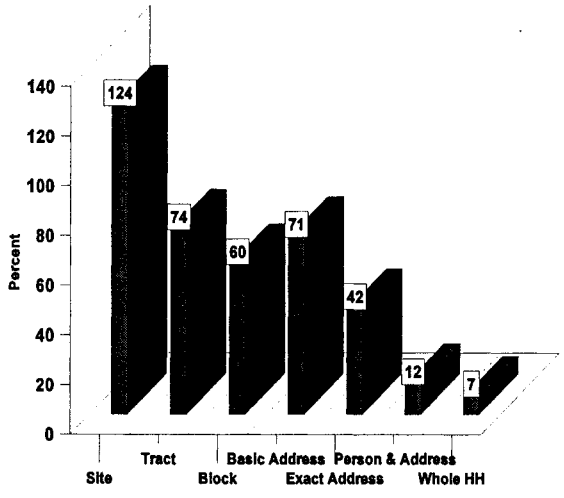
This evaluation of the 1996 Administrative Records Database indicates that it cannot yet provide the basic information about the population obtainable from a population census.

There is evidence of measurement error from the evaluations. Here we reexamine our results to highlight areas where more research and development (R&D) is needed to assure better quality databases and evaluations. This constitutes a research agenda of some of the things to be learned, tried and solved before administrative records become a viable alternative to primary data collection.

We have used many different indicators of the discrepancy between ARD and census counts or match rates between ARD and census records.

For this discussion, we have taken the liberty to plot all of them on one axis in Figure 1. Values close to 100 are desirable, values close to zero are not desirable. As values move above 100, they become increasingly undesirable.

Figure 1. Comparisons between ARD and Census Counts



At the far left of the figure, the overall ratio of ARD count of people to the census count is 1.24 ((20,534 + 4,869) / 20,495) at the site level. That is, there are 124 ARD people in the database for every 100 people on the Census list of people. This is a substantial reduction in bias compared to the Phase 1 databases for the 1995 Test Census sites which had site ratios of about 3.0. But there are still too many ARD person records in the database. The ARD may contain people who have moved or died and it may contain more than one record for some people. Future R&D should explore several things:

(1) the ability of a national database to unduplicate people records in order to solve the mover problem,
(2) increasing the power of the record linkage algorithms to unduplicate records from several sources (recognize the same person when contributed by different sources),
(3) adding new source files to identify name changes and other personal changes that hamper unduplication,
(4) treating people in special places appropriately and consistently when comparing the ARD and census,
(5) the ability of additional files (e.g., National Death Index, Emigration Records, and Special Place enumerations similar to current census procedures) to assist to identify transitions.

Our MARE index portrays a disagreement rate. For discussion, here, we define 1-MARE as an agreement rate. Then there is about a 74 (60) percent agreement on population size between administrative records and the census at the tract (block) estimation level. This may represent an upper bound on the percentage of people whom we could match between census and administrative records if we could solve the major record linkage problems mentioned below. Assigning people to correct blocks and tracts depends on having good address information from ARD sources, and being able to "geocode" it accurately to small geographic areas. Part of the problem is that our current geocoding procedures cannot assign P.O. Box addresses to geographic areas. Such addresses are not uncommon in administrative records (see below). Therefore,

(6) Further R&D on procedures to geocode the kinds of addresses found in administrative records is needed.

We can match 71 percent of the ARD address records to census address records at the level of the basic street address. One major reason for failing to match addresses is that 27 percent of the administrative records addresses were P.O. Box mailing addresses instead of the physical location addresses that were used in the census. An important part of future R&D activities for developing administrative record databases will be:

(7) developing ways of converting mailing to physical location addresses.

The National Academy of Sciences Panel has suggested acquiring and using local records that contain both kinds of addresses such as property tax rolls and utility company records. This would add considerably to the processing effort for a national database; such local acquisition and processing might be handled by state data centers to ease the burden on the central facility.

Only about 42 percent of the ARD addresses could be matched exactly to census addresses (including a match on the

unit designator, if applicable). Obviously, linkage problems arise because of the absence or the inconsistency of unit designations on administrative records. If it is necessary to assign a person to an exact residential unit in an administrative records database, then:

(8) a solution must be found for the missing or inconsistent unit designator problem in administrative records addresses.

Research might investigate various modeling alternatives to create synthetic families that closely approximate observed families within basic street addresses. Or one might investigate treating multifamily addresses as special places and request a list of occupants from the building manager for Census Day. Multifamily buildings and their managers would have to be identified ahead of Census Day for this procedure to work. The commercial lists might be useful.

Address matches might be improved by changing the record linking strategy. We matched each administrative record to the geographic database on address to obtain an identification number. That unique, identifying number represents the geographical hierarchy of information about each address (its state, county, tract, block, street name, street number, etc.). Then records were linked using this number. A procedure that allows probabilistic linkage across records on the standardized components of the address probably would result in many more linkages (some of which would be false). In more general terms, we need R&D:

(9) to develop optimal strategies for using existing record linkage procedures to match addresses.

We matched 12 percent of the ARD people to the same census people at the same census address. This is a very low rate. Again, procedural changes may produce big gains if, for example, people linkages unconditional on address are allowed. Of course, if address information can be improved and more address matches made (see suggestions, above) then more people matches conditional on address matches can be made also. But there are still many people in the Census lists that could not be found in the ARD. So, more people records are needed in the ARD and the implication for the future research is to:

(10) investigate ways of including more people who are otherwise found only on the census.

(11) investigate the ability of longitudinal administrative record databases of people to provide more accurate information about a current population.

Finally, we defined households according to census information and searched for groups of administrative records people that exactly matched the defined census households. Only 7 percent of the administrative records people could be linked to whole census households using these criteria. So, it would not be feasible to use this administrative records database as a direct source of accurate information about households that did not respond with a census form.

However, more sophisticated methods e.g., modeling approaches, have been suggested by others. It will be important to continue research:

(12) to find methods of using administrative records information from same and neighboring areas to impute values for missing items on census and survey forms.

Although not shown in the figure, there were consistent results for the quality of ARD information about personal characteristics: several analyses showed that, for matched people, the agreement rates were good for race and Hispanic origin (ignoring missing data) but not as good when taking missing values into account. So research is needed:

(13) to improve, by using lists of names associated with particular demographic characteristics, the quality of information about race and Hispanic origin inferred from administrative records.

(14) to use additional administrative records sources to obtain multiple observations about each person. Use the additional information in measurement models to improve the quality of the measures of characteristics such as race, and Hispanic origin.

A final observation: There appear to be many more black people in the ARD than in the census at the block and tract levels. Several advisory groups have questioned whether minority groups are well represented in administrative records. This evaluation, as far as it goes, shows promise. Future evaluations should address:

(15) how well do administrative records represent minority groups.

V. References

1) Buser, P., Huang, E.T., Kim, J and Marquis, K. (1998). 1996 Community Census Administrative Records File Evaluation. 1996 Community Census Memorandum Series Number 20. From Tommy Wright to John H. Thompson.

2) Winkler, W.E. (1994), "Advanced Methods for Record Linkage, American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 467-472.

3) Winkler, W.E. (1995), "Matching and Record Linkage," in B.G. Cox (ed.) Business Survey Methods, New York: J. Wiley & Sons, pp. 355-384.

The authors would like to thank Vickie Kee, Jeong Kim and Diane Simmons for records matching work; Pascal Buser for generating tables; Charlene Leggieri, Florence Abramson and Arona Pistiner for comments on the earlier version of the paper. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than the official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.