RECORD LINKAGE FOR THE 1982 CENSUS OF AGRICULTURE MAIL LIST DEVELOPMENT USING MULTIPLE SOURCES

Jane Y. Dea, Tommy W. Gaulden, and D. Dean Prochaska, U.S. Bureau of the Census

1. INTRODUCTION

The development of a mail list is one of the most significant phases of the overall task of taking the census of agriculture, since accurate census results are highly dependent upon a complete mail list. The objective in development of the list is to obtain a complete list while minimizing duplicate records and eliminating nonfarm records. The mail list for the 1982 Census of Agriculture is a prime example of a single-use census list compiled from multiple administrative record sources. A three-part automated record linkage system was developed as a cost-effective means to identify duplication: (1) Employer Identification Number and Social Security Number linkage, (2) name/address recode linkage, and (3) clerical review encompassing all record sets not previously defined as duplicates or nonduplicates. The alphabetic linkage part of the system was based on the record linkage theory developed by Fellegi and Sunter of Statistics Canada.[1] This paper describes and analyzes the methodology used to remove duplicates and identify nonfarms from the multiple source list through a record linkage process.[2]

2. CENSUS BACKGROUND

The 1982 Census of Agriculture was the 22nd nationwide census of agriculture taken in the United States. The first agriculture census was taken in 1840 in combination with the census of population. The census is required by law under the provisions of Title 13, U.S. Code and generally is taken every 5 years in the 50 states, plus Puerto Rico, Guam, and the Virgin Islands. The 1982 census and the three previous censuses were conducted primarily by mail with data collection by self enumeration. Prior to the 1969 Census of Agriculture, data were collected by enumerators through personal interview. The initiation of census data collection by mail required development of new procedures and methods for procurement of administrative record source lists and handling of large name and address files in the list development, mailing, and processing operations.

Census report forms are mailed at the end of the census reference year with follow-up letters and report forms being sent to nonrespondents at 3- to 4-week intervals.[3] Data collection requires approximately 6 months. Telephone follow-up is used to obtain data for nonrespondents thought to have large operations. After the data collection phase, report forms are checked and processed primarily using computer assisted methods. The final processing operations include data table preparation, technical review, and publication of the census data.

3. ADMINISTRATIVE RECORDS

The availability and procurement of administrative record files are major requirements for collection of data by mail for the census of agriculture. Early research studies indicated there was no single source file which would provide adequate coverage for the census. Therefore, a combination of several different administrative record files was used in order to obtain as complete a list as possible.

The primary source lists used for the 1982 Census of Agriculture were the files of farm operators from the previous 1978 Census of Agriculture, the Internal Revenue Service (IRS) file of individuals filing Form 1040 Schedules F or C (farm tax returns), and the producers files of the Agriculture Stabilization and Conservation Service (ASCS) of the U.S. Department of Agriculture (USDA). Other source lists used included the IRS farm partnership file (form 1065), the IRS farm corporation file (form 1120), the Social Security file of farm employers (form 941 and/or 943), the USDA Statistical Reporting Service (SRS) list frame file for the 31 available states, the nonrespondent file from the previous census, and special lists from various sources for large or specialized farm operations. In addition, the nonfarm records and the duplicate records from the previous census were used to facilitate farm status classification and removal of duplicates. The information obtained for the source lists from outside the Census Bureau was limited and varied by source. A majority of the lists had some type of code or value which indicated the size of operation while others had farm location and type of operation indicators. The quality and up-to-dateness varied by source.

The total number of name and address records obtained from all sources was about 19.0 million. There was extensive duplication between files and within files. Variations of the same name-nicknames, initials, middle names, and farm names appeared in the source lists. Farm operators used different addresses due to business and residential locations or relocations.

4. METHODOLOGY

The development of the census mail list consisted of two list building phases: (1) the Farm and Ranch Identification Survey phase that included 15.8 million source records, and (2) the main census phase that provided an additional 3.2 million source records. Each phase had five major operational parts: (1) Format and Standardization, (2) Employer Identification Number (EIN) and Social Security Number (SSN) linkage, (3) Geographic coding and ZIP Code edit, (4) Alphabetic name linkage, and (5) Clerical review of all record sets not previously defined.

The first phase was completed in early 1982 and a preliminary list resulted. Units identified from this phase as having a high likelihood of being nonfarm (based primarily upon the list source or combination of sources) were selected for inclusion in the Farm and Ranch Identification Survey. The objective of this survey was to identify nonfarm addresses and add new tenant and successor names. The results of the survey, along with previously unavailable

source lists, were used in the second phase of record linkage to develop the final census mailing list. This two-phase process reduced the list from 19.0 million addresses to 3.6 million addresses. Quality control samples and a "trace" sample were used during production processing to validate methodology, provide mail file estimates, and test computer programs.

4.1 Initial Processing

Before multiple source records could be linked and duplicate records eliminated, the individual source records needed to be put into a standard format for name and street or rural address. As part of the process of providing name and address format and standardization, many operations were performed to provide tools for the subsequent record linkage and duplicate identification. These included an edit of the source record, a determination of name control, the insertion of a surname locator, the identification of address components, the assignment of a size code, the identification of a potential partnership or corporation with a record flag, and geographic coding.

The basic edit program placed all source records into a common format for processing. The format used consisted of four types of fields: (1) primary and secondary name field, (2) address field, (3) place field (city, state, and ZIP Code), and (4) processing code fields. Each record was assigned an address priority code to identify the source list. This code was used in the linkage process to determine which source record to retain in the case of duplicates. Source lists with the surname first were edited using a program to switch the order of names.

The edit program also removed commas, periods, and certain special symbols from the name and address fields and inserted a space between any adjacent numerics and alphabetics. For example:

James F. Jones, Jr.	became	James F Jones Jr
1420 Elm #301	became	1420 Elm 301
76B598	became	76 B 598

By this process, the name and address fields were broken down into a series of numeric or nonnumeric words separated by one space.

Name control (normally the first four characters of the surname) was essential in determining positive or possible duplicate status when records were linked on EIN or SSN. Although name control existed on many source records, the various sources used different procedures. A uniform method was designed and used on all records to identify the surname. This program involved reading the name field and matching selected words to a "skip list" dictionary containing over 1,000 words and abbreviations (such as Farm, Dairy, Bros) which could appear in the name field, but were not likely to be the surname. An indicator (surname locator) was placed in each record to identify the field position of the word used to derive the name control. It was used later in identifying name parts for recoding.

Numeric characters were extracted from the address field for use in determining match status in the alphabetic name linkage. Box numbers, rural route numbers, and street address numbers were identified and placed in specific data fields. One field contained box numbers and street address numbers; a separate field contained rural route numbers. A subroutine of the edit program was designed to scan the address field for numeric words and classify them according to their position relative to nonnumeric words matched to a dictionary. The address completeness and characteristics varied considerably by source--20.4 percent contained Box or Street and Route, 33.1 percent had Box or Street only, 28.1 percent had Route only, and 18.4 percent had neither.

Each record was assigned a measure of size derived from size indicators present in the source record. A size code was placed in a separate field for each source. During record linkage, the size code was retained for all sources on which a name appeared by transferring data from the deleted duplicate record to the retained record. This allowed the derivation of both a "source combination code" indicating all the sources for the final record, and a "final size code" from all the individual source size codes. The final size code was used in census processing to determine the type of report form to mail, sampling rate, and type of follow-up procedure for nonrespondents.

The record linkage process was designed to prevent computer deletion of matched partnership or corporate records and individual records. Because individuals are commonly involved in both partnership and sole proprietorship operations, records that possibly represented a partnership or corporation were flagged. This flag (known as the PPC flag) prevented erroneous computer deletion of records with matched names or identification numbers, permitting a clerical decision to be made on the linked records. A dictionary of words and abbreviations associated with partnerships and corporations was used as a basis for applying the PPC flag to a record.

An essential part of the processing was geographic coding. The geographic coding system was designed to ensure that each of the records entering the record linkage system contained standardized and edited geographical codes; i.e., state and county codes, county alpha codes, ZIP Codes, and ZIP group numbers. Numerical state and county codes were assigned based on ZIP Code. County alpha codes consist of the first four letters of the county name and are used in census processing. For the majority of records the ZIP group was identical to the 5-digit ZIP Code. However, in the cities served by multiple ZIP Codes a single ZIP group number was assigned for all ZIP Codes in the city range, thus treating the city ZIP range as a single ZIP Code.

4.2 Record Linkage

The actual record linkage was done in three stages--linkage by Employer Identification Number (EIN) and Social Security Number (SSN), linkage by name within geographic ZIP groups, and linkage with historical agricultural records. The most direct means for linking records among the source lists was by the use of the EIN or SSN. Eighty-nine percent of the records had either EIN or SSN or both. After linkage on these identification numbers and deletion of positive duplicates,

names and addresses were recoded prior to performing alphabetic name linkage within a ZIP group number (or block). At the completion of this stage each linked record was classified as a duplicate, possible duplicate, or nonduplicate. Possible duplicates were reviewed clerically. The details of the operations performed in these three stages of record linkage are described in this section.

The identification number record linkage stage had a separate process for EIN linkage and for SSN linkage. Records with both numbers were carried through both matching operations. The first step in the operation was to sort the records by EIN or SSN, name control, PPC flag, and address priority. The sort was a critical factor for the proper functioning of the system since several variables were checked within a block of records with the same EIN or SSN. If records matched on EIN or SSN, but the name control variable was not equal or the PPC flag was present, the records were identified as possible duplicates for clerical review.

Most of the records from the Form 1040 Schedules F or C contained two SSN's -- usually husband's and wife's. Since the linkage was accomplished by sorting individual records by SSN, "dummy" records were created for those records with two SSN's. The "dummy" records were the exact duplicates of their masters except the SSN's were reversed, thus allowing linkage on both numbers. After the linkage process, the "dummy" records were matched back to their master records and any codes picked up during processing were transferred to the master. The "dummy" records then were deleted.

When two records were linked a comparison of other record variables was made. On this basis, the records were identified as either positive duplicates or possible duplicates. In the first case one of the linked records was deleted by the computer. In the second case, the linked records were displayed and reviewed clerically. Additional information in the records was used to determine match status according to specific rules and procedures. When positive duplicate status resulted, the objective was to retain the record with the highest quality address and information (e.g. 1978 census codes, if present, Standard Industrial Classification Code, geographic codes, and the source size indicator) from the record to be deleted by transferring it to the retained record. The PPC flag was used to change the match status to possible duplicate when possible partnership or corporate type names were involved.

In order to perform alphabetic name linkage, names and addresses had to be put into a standard format. For names, this involved identification of each name part of a record, creation of alternative multiple name patterns, coding of surname and first name, conversion of nicknames to proper names, and recoding of those names. Address recoding was completed in the initial processing.

In order to identify each name part of the individual records, each word in the name field was compared to the "skip list" dictionary. Those words appearing on the "skip list" were deleted. All remaining words were classified as either a surname, single letter, conjunction, or other. The surname was identified by the surname locator. Nicknames and conjunctions were identified through "look-up" dictionaries. Words and letters were classified by codes given in the following table. These codes were retained in sequence and became the name pattern.

Word Classification and Coding

Word Type	Code
Single Letter	2
Surname	3
Conjunction	4
All Others (including nicknames)	1

The name pattern was compared to a name pattern file which identified each word or letter as a first name (FN), first initial (FI), middle initial (MI), or last name (LN). When multiple name patterns were encountered, additional output records were created. Multiple names were identified as names following a conjunction such as "&," "and," "or," etc. Additional output records were created for names in the second name field and partnership names.

If the character following the middle name is a conjunction, and the name pattern is "John Jones & Frank Small"--pattern = 11413, then three names were recoded--John Jones Small, John Jones, and Frank Small. Note that this pattern also recodes "John Paul & Mary Jones" into John Paul Jones, John Paul, and Mary Jones. This is an attempt to identify partnerships which could change name order in different source file records. Examples of name pattern recodes are given below.

Example (1):

Name	Robert	E	Patterson
Name Pattern	1	2	3
Word Type	Other	Single Letter	Surname

Recode: FN=RBRT
FI=R
MI=E
LN=PTRS

Example (2):

Name	John	A	&	Mary	C	Doe
Name Pattern	1	2	4	1	2	3
Word Type	Other	Single Letter	Conjunction	Other	Single Letter	Surname

Recode 1: FN=JHN
FI=J
MI=A
LN=D

Recode 2: FN=MR
FI=M
MI=C
LN=D

Example (3):

Name	John	Jones	&	Frank	Small
Name Pattern	1	1	4	1	3
Word Type	Other	Other	Conjunc-tion	Other	Surname

Recode 1:	Recode 2:	Recode 3:
FN=JHN	FN=JHN	FN=FRNK
FI=J	FI=J	FI=F
MI=J	MI(none)	MI(none)
LN=SML	LN=JNS	LN=SML

In all phases of alphabetic name linkage about 99.3 percent of the name and address input records were matched to the pattern file and 0.7 percent were rejected as nonpattern arrangements. A nonpattern arrangement occurred when the surname locator was blank or when a particular pattern did not match one of the possible name patterns. This occurred primarily in multiple name strings, such as "Tom A Dick B and Harry C Smith."

After identification of all patterns of name parts, surnames and first names were recoded using a soundex system modified previously for use in the agriculture census. The recoded name retained the first letter, deleted the second of each double consonant and all vowels including Y, and truncated the name to four characters. Thus, the name DILLINGER was recoded as DLNG.

Nicknames such as DICK, BILL, BECKY were converted and had their proper names RICHARD, WILLIAM, REBECCA recoded instead, in order to standardize different versions of the names used on different source lists. Also, abbreviated versions such as ED, GEO, WM were converted and had their proper names EDWARD, GEORGE, WILLIAM recoded. This was accomplished through a match of the first name to a "Nickname Dictionary."

Alphabetic name linkage was then performed on the recoded file to identify duplicate records. Linkage was attempted within a limited specified group of records or "block"--a 5-digit ZIP Code or ZIP group number. All records were merged and sorted on recoded name and address within each "block." The records were then compared in a pairwise fashion based on their sorted order within the linkage "block."

Each linked record was classified as duplicate, possible duplicate, or nonduplicate. It was desirable to classify and eliminate by computer as many duplicates as possible, yet retain names which represent separate agriculture operations. Six match variables (last name, first initial, first name, box/street, route number, middle initial) were used to classify the name records. Last name (LN) and first initial (FI) were required matches before further comparisons were made on the remaining variables. The comparisons were made on all combinations of variables and classification was based upon the presence and extent of agreement between the match variables. All pairwise comparisons were made for adjoining records with the same LN and FI, such that the maximum number of comparisons was ${}_nC_r$ where r = 2 and n = the number of records having the same LN and FI.

Certain combinations of variables were given greater importance in determining the classification on the basis of the uniqueness of the combinations. The classification of each combination was determined primarily on validation checks of sampled linked records throughout the processing operation and from previous censuses. Based on the importance given each possible combination, a set of numerical weights was developed by an independent group in order to test the consistency of the match classification of each combination. The result of the analysis demonstrated that, with minor exceptions, the match classifications of each combination are consistent based on the underlying combination importance assumptions.

When each match variable was compared, one of three data classifications resulted:

= & > 0 The match key is equal in the comparison set and not blank. (Data not conflicting.)

≠ & > 0 The match key is not equal in the comparison set and not blank. (Conflicting data.)

= 0 The match key may be blank in both records in the comparison set, or may be present in one record but not in the other. (Data cannot be compared.)

Duplicate records matched on both first and last names as well as address information. However, if one of the records in the set had Jr. or Sr. attached to the name, the match status was changed to possible duplicate and displayed for clerical resolution. But, if the two records had conflicting Jr. or Sr. names, these records became nonduplicates. Possible duplicate records matched on the first and last name, but address information was not present or did not match. Records with the first initial only that matched on the last name and address also were included in the possible duplicate group. Nonduplicate records did not match on last name recode. Records with the same last name recode but with different first initials also were included in this group. In most instances, records where the middle initial did not match were included in this group.

Example 1: Classified as a duplicate for deletion by computer

Record 1:	Record 2:
John A Doe	John A Doe
Box 123	Rt 4 Box 123
Suitland, MD 20233	Suitland, MD 20233
Combination #8	
FN = & > 0	RR = 0
Box = & > 0	MI = & > 0

Example 2: Classified as a possible duplicate for clerical review

Record 1:	Record 2:
A B Smith Jr	A B Smith
Rt 2 Box 34	Rt 2 Box 34
Hyattsville, MD 20784	Hyattsville, MD 20784

Combination #29

FN = 0 Box = & > 0 MI = & > 0
FI = & > 0 RR = & > 0

Example 3: Classified as a possible duplicate for clerical review

Record 1:	Record 2:
John A Smith	J A Smith
Route 2	Rt 1
Goose Lake, IA 52750	Goose Lake, IA 52750

Combination #50

FN = 0 Box = 0 MI = & > 0
FI = & > 0 RR ≠ & > 0

Example 4: Classified as a nonduplicate

Record 1:	Record 2:
John A Smith	John B Smith
RR 1	RR 1
Goose Lake, IA 52750	Goose Lake, IA 52750

Combination #21

FN = & > 0 RR = & > 0
Box = 0 MI ≠ & > 0

When the recoded name records were linked and classified as duplicates, data were transferred from lowest priority address to highest priority address before deletion. When a possible duplicate was identified, no data were transferred and the data sets were displayed for clerical resolution. The clerks compared the linked pairs, determined match status, and when records matched, determined which record(s) to delete as described in EIN/SSN linkage. Linked records classified as nonduplicates received no action and were retained as separate records in the file.

After completion of the EIN/SSN and alphabetic name linkages, an additional linkage process was performed using historical information. This was an additional clerical review which included multiple record sets identified in the previous census of agriculture and their associated linkages from the EIN/SSN and the alphabetic name linkage processing. The records in these sets usually contained no common names. The additional linkage process enabled these records to be sorted together for review. Sets including a partnership or corporate record were displayed and considered for inclusion in the Farm and Ranch Identification Survey for identification of duplication by respondents.

4.3 Farm and Ranch Identification Survey

Completion of the first phase of record linkage resulted in a file of approximately 7.3 million records, including some nonfarm source records. Each record was classified into one of three groups based primarily on source or combination of sources:

(1) Probable Nonfarm (2.3 million)--Nonfarm records from the previous census which failed to match any other record or matched to certain single sources only.

(2) Probable Farm (1.9 million)--Multiple-source records usually including a match to a 1978 census farm.

(3) Farm Status Questionable (3.1 million) --Nonfarm records matching to other sources, records not matching a 1978 census farm, and certain single-source records.

The "Probable Nonfarm" group was deleted, resulting in a preliminary mail file of 4,969,809 records. Of this total, 3.1 million records with source and size code (most likely to represent nonfarms) were selected for inclusion in the 1982 Farm and Ranch Identification Survey. Approximately 50,000 or 2 percent of the mail list were included to resolve potential duplication between individual name records linked to possible farm associated businesses. These cases were mailed a short, one-page report form designed to determine whether an operator qualified as a farm and, if so, the approximate value of its sales. Respondents also were asked to provide the names and addresses of any tenants or succeeding operators. During survey processing, these names were searched on the preliminary mail file and, if not matched, were added as another source list in the second phase of record linkage.

The survey was mailed in early March 1982 and included a series of follow-up mailings to nonrespondents over the next several months. The information obtained from this survey was used to update the addresses in the preliminary mail file and provide information on farm and nonfarm status in preparation for development of the census mail list. This survey yielded 2.5 million receipts--response rate of 82.9 percent. The receipts identified 1.2 million nonfarms and 816,000 farms. Records resulting from the tenant and successor search totaled 38,840.

4.4 Final Processing and Results

The record linkage system for the main census list phase was similar to the one used for the farm and ranch survey phase. Several new source files consisting of 3.2 million names were added. These source files included the IRS files for tax year 1981 (1040 F and C filers, 1065 partnership file, 1120 corporation file, and 941 and/or 943 farm employer file), tenant and successor adds from the farm and ranch survey, and some additional special lists. These new source records and the records from the farm and ranch phase were processed through the same record linkage system with some modifications. The 3.2 million new source records produced 413,000 additional addresses after the linkage process. After completion of all phases of linkage, the final mail file of 3.6 million records resulted.

5. EFFECTIVENESS OF THE LINKAGE SYSTEM

The objective of the record linkage process is to develop a mail list in a cost-effective manner from multiple sources that cover the universe of agricultural operations. The methodology used in the process is effective in reducing the cost of the census data collection operation to the extent that it succeeds in identifying and eliminating duplicate farm operations and nonfarm records from the list. Census coverage suffers though if qualifying operations are eliminated at the mail list development stage. The list development process needs to balance these concerns as well as have a relatively low cost due to the large volume of records.

An estimate of about 150,000 duplicate records remaining on the final census mail list was developed from counts of duplicates identified by respondents and in census processing. This number is approximately 4 percent of the 1982 mail list and compares with about 11 percent of such duplicates on the 1978 mail list. When census processing is completed, a more accurate estimate of mail list duplication will be made.

Using nonfarm records from the 1978 census in the record linkage operation was effective in eliminating records not qualifying as farms. This technique made it possible to significantly reduce the final size of the mail list without risking a significant loss in census coverage. The 3.6 million mail list records for 1982 compared with 4.4 million in 1978, a reduction of almost 20 percent. An evaluation of the coverage of the 1982 mail list is presented in Coverage Evaluation for the 1982 Census of Agriculture[4]. This evaluation estimates some types of error from the record linkage process as well as from other sources.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the assistance of Margaret Bruce, Eileen Gray, and Kathleen Campbell for their invaluable manuscript word processing, and a number of colleagues for their constructive comments.

NOTES AND REFERENCES

[1] Fellegi, I.P. and A.B. Sunter, "A Theory for Record Linkage," Journal of the American Statistical Association, 1183-1210, December 1969.

[2] For more detailed documentation of the record linkage system, the authors may be contacted at:

U.S. Bureau of the Census
Agriculture Division
Program Research and Development Branch
Washington, D.C. 20233

[3] Ruggles, Donna R., Jane Y. Dea, Flora K. Kwok, and Cindy A. Carman, "Evaluation of the Effectiveness of Data Collection Procedures for the 1982 Census of Agriculture," 1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.

[4] Davie, William C., Emily Lorenzen, and D. Dean Prochaska, "Coverage Evaluation for the 1982 Census of Agriculture," 1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.