CONTENTS

Chapter 3. Preparatory Operations

INTRODUCTION

Discussions about refining plans for the 1987 agriculture census continued practically up to the date of the first census mailing, but by early 1986 specific preparations for the enumeration itself began. These preparations included three major activities:

1. Compiling the census mail list
2. Printing and addressing the report forms
3. Promoting the census (for information on the promotion campaign, see ch. 4)

In addition, continuing discussions with the Office of Management and Budget (OMB) about the size and composition of the initial mailout resulted in the design and test of a new "short" screening report form for use in the census mailing.

Preparing the census mail list required the acquisition, compilation, and "linkage" of records from various sources. Linkage involved identifying duplicate records, analyzing source and address codes, and deleting the record with the lower priority (i.e., from sources considered less reliable in providing current and complete mailing addresses) codes after transferring the identification data to the higher priority record, and attempting to classify by size the various addresses believed to represent farms. A linkage operation was carried out as part of each of the two major address list assembly operations in the winter and spring of 1986-87, and the summer and fall of 1987. After each linkage operation, the Bureau applied a classification tree model (see p. 28) to the resulting address lists, grouping addresses according to their likelihood of being a farm. Once the final list was compiled, final preparations included assigning census file numbers and sampling for data collection.

The Bureau finalized the content of the standard report forms in January 1987, and tested a "screening" form in December 1986 and early 1987. The screening form (87-A0400) design was finalized in March 1987. (See app. F for a description of the report forms and for facsimiles of representative forms.) The report forms and other enumeration materials were printed by private contractors who also assembled the bulk of the mailing packages. Once the mailing list for the first mailout was complete, the Bureau's Jeffersonville, IN, facility printed the address labels and affixed them to the mailing packages. Approximately 4.1 million mailing packages were prepared for the first census mailout in December 1987.

ADDRESS LIST DESIGN

Introduction

The Census Bureau introduced the mailout/mailback procedure for the agriculture census in the 1969 enumeration. While more economical in both workforce and funding requirements than the personal interview enumeration, the mail census requires a complete and accurate list of addresses for operations that meet the census definition of a farm. Moreover, the ideal list should not include duplicate addresses, or addresses that do not meet the census farm definition; every duplicate or nonfarm address means additional mailing costs, and added response burden that can undermine the cooperation of respondents and the accuracy of the data collected.

Thus, compiling the mail list is a major part of the census operation. This was particularly true for the 1987 enumeration because of the limit set on the total size of the list, and the number of standard and sample report forms that could be included in the first mailing. The OMB directed the Bureau to restrict the number of packages in the first census mailing to no more than 4.2 million; approximately 3.2 million could be regular and sample report forms, while the remainder were to be screening forms used to determine whether suspect addresses met the census criteria for farms. The 1982 census mailing had been somewhat smaller (3.65 million packages), but had been preceded by a Farm and Ranch Identification Survey mailing to over 3.1 million addresses, and the results of that mailing were used to reduce the total size of the census list.

General Procedures

The Bureau compiled the mail list for the 1987 Census of Agriculture from previous census mail lists and from current or nearly current administrative records from various Federal agencies concerned with agriculture. The list was assembled in a two-phase operation, the first between October 1986 and April 1987, using records from previous censuses and the latest available administrative records, and a second phase, carried out between July and November 1987, with additional addresses drawn from the National Agricultural Statistics Service (NASS) and the Internal Revenue Service (IRS). This enabled the Bureau to (1) include more recent records in the second phase than were available for the earlier operation, (2) refine procedures for greater efficiency, and (3) review the use of the classification tree model employed to identify groups of addresses according to the expected proportion of census farms in each group.

The two processing phases employed similar procedures, and included seven major operations:

1. Uniform record formatting
2. Matching and deleting ("linking") duplicate records based on employer identification number (EIN) or Social Security number (SSN)
3. Geographic coding
4. Matching and deleting duplicate records based on name and address
5. Assigning source and size codes
6. Manually reviewing possible duplicates
7. Identifying groups of records by expected (or estimated) proportion of census farms in each group

Sources

Preliminary list—The Bureau began preparing the 1987 mailing list in October 1986, using the main computer facilities at Suitland, MD to compile and process the preliminary source list records. This first-phase linkage operation involved approximately 10.2 million records, drawn from the following sources:

Source	Records
Total	10,242,159
1982 Census of Agriculture farm list	2,027,123
1982 Census of Agriculture non-respondents	563,848
1982 Census of Agriculture nonfarms	986,360
1982 Farm and Ranch Identification Survey nonfarms	1,258,275
NASS farms	1,759,378
NASS nonfarms	488,457
1985 IRS, SSA records*	3,120,644
Special commodity lists	38,074

*Included IRS records for the following forms:
1040F Schedule for Farm Income and Expenses (attached form 1040, Individual Tax Return)
1120 Corporation Income Tax Return (for SIC codes 01, 02, 07)
1065 Partnership Return of Income (for SIC codes 01, 02, 07)
941/94 Employers' Annual Tax Returns for Employers (941 coded SIC 01, 02, and 07 (Agriculture) for nonagricultural workers, and 943 for agricultural workers)

Final list—The Bureau completed the first linkage operation in April 1987 and produced a preliminary mail file of 5,921,660 addresses. All of these addresses were included in the final linkage process—begun in July 1987—together with approximately 3.2 million additional records not available for the first-phase processing. Altogether the second phase mail list linkage program included over 9.1 million records, as follows:

Source	Records
Total	9,174,143
Preliminary list records	5,921,660
1982 Census of Agriculture farms	1,385
NASS farm adds	209,161
NASS nonfarm adds	623
1986 IRS, SSA records	2,911,840
Special commodity lists	113,595
USDA June Enumerative Survey (JES)	15,879

The second-phase matching and linkage operation was completed in September 1987, at which time the mail list consisted of 6,043,157 records.

Source priority codes—The Bureau assigned an address source priority code to each source providing records for the agriculture census mail list compilation operation. These codes were assigned based on the probability of a given source file containing a high percentage of complete and current addresses. (The determination itself was based largely on experience in mail list compilation from past censuses.) The principal sources and their priority codes were as follows:

Source	Priority code
Multiunits and abnormals*	1
IRS 1040F list	2
IRS 1065 list	3
IRS 1120 and 1120S list	4
IRS 941 and 943 list	5
1982 Census of Agriculture inscope list	6
NASS list frame	7
Special lists	8
1982 Census of Agriculture non-respondents list	9
1982 Census of Agriculture out-of-scope list	10
1982 Farm and Ranch Identification Survey out-of-scope list	11
NASS list-frame nonfarms	12

*"Multiunits" were companies or organizations with substantial agricultural operations in more than one location. "Abnormals" were farms operated by institutions, such as State research facilities.

Format and Standardization

General information—The Bureau collected over 12 million separate names and addresses from various sources in the two-phase mail list compilation operation, many of which were duplicates. Before any linkage could identify and delete duplicates, the agency had to establish a computer record format compatible with its processing programs, and standardize the different computerized records assembled from the source lists. This required a series of operations to identify each record's components and to reformat them as necessary, including (1) assigning a source file number (SFN—a unique identification number) to each record; (2) editing each record; (3) determining name control (usually the first four surname letters); (4) inserting a surname locator; (5) identifying each address; (6) assigning size codes; and (7) assigning potential partnership or corporation (PPC) flags.

In addition, the format and standardization program identified agricultural services records and records for ZIP Codes outside the 50 States for deletion from the mail file. The following computer files were established to help computer processing of the mail file:

Agriculture services records (for deletion)

Records with ZIP Codes outside the 50 States (for deletion)

Trace sample (used to evaluate processing efficiency)

Records without an employer identification number (EIN) or Social Security number (SSN)

Records with EIN's and/or SSN's

"Short" records with EIN's with or without SSN's

"Short" records with SSN's only

"Short" post office name records for records without EIN's or SSN's

Tally file (by size and geographic code)

All serialized records (all records from the input files with their source file numbers (SFN's)

Source file numbers (SFN's)—The format program assigned a unique identification number to each computerized record to locate and identify that specific record, and the source from which it was drawn. Ranges of eight-digit numbers were reserved for each source used in the mail compilation and the computer program assigned numbers from these ranges to the appropriate records during the initial processing run.

Source record edit—The basic edit program placed all source records (i.e., names and addresses from the various sources) into a common format for computer processing. The common format used consisted of four types of fields:

1. Primary and secondary name
2. Address
3. Place (city, State, and ZIP Code)
4. Processing codes

The edit program assigned an address priority code to each record to identify the specific source list of origin for use in the linkage operation, and thus determine which source record to retain in the case of duplicates. A special program was used to edit source lists with surnames first, switching the order of the names (e.g., changing "Smith, John," to "John Smith").

The primary edit program also removed commas, periods, and certain special symbols from the name and address fields, inserted spaces between adjacent numeric and alpha characters, and substituted standard two-digit State abbreviations for State names or old-style abbreviations. For example:

Mr. James M. Doe, Jr.		Mr James M Doe Jr
2429 State #345	became	2429 State 345
Hodag, Wis. 55555		Hodag WI 55555

Name control—The "name control" for a record usually consisted of the first four characters of the surname, and was used to determine possible duplicate status when linking records based on the EIN's or SSN's. Many of the source records used for the census list already had name controls, but the various sources used different methods for assigning them; before linkage could proceed, the Bureau had to establish a uniform method to use on all the records. The formatting program read the name field in each record from right to left until it identified a nonnumeric word with three or more characters, and matched that word to a "skip list" dictionary containing a list of words and abbreviations (such as "Farm," "Dairy," "Bros.," "& Sons," and so on) that might appear in the name field but were unlikely to be the surname. The first nonnumeric word with three characters or more that was not on the "skip list" was used to determine the name control for that record. The first four characters (from left to right) of this word were inserted in the name control field. If the computer program found no usable word after scanning the entire primary name field, the original name control provided was used. (If none was provided, the field was left blank.)

Surname locator—The formatting program inserted an indicator—the surname locator—in each record to identify the field position of the first character of the name control. If the name control field was blank, the record could not be recoded (for details of the name recode, see below) for name and address linkage.

Address identification—The formatting program organized the source records' addresses for alphabetic name linkage. Numeric characters were identified and extracted from the address field. The program identified box, rural route, and street address numbers and placed them in two specific data fields, one for rural route numbers, the other for box and street numbers. The program scanned each address from left to right until a numeric "word" (i.e., one or more numeric characters) was identified. If it was the first word in the field, it was stored in the box/street field; otherwise the word preceding the numeric word was matched to a dictionary of acceptable words (e.g., "Box," "RFD," "Rte,"—although words such as "No." or "Number" were ignored). If the word matched one of those in the dictionary, the numeric then was stored in the appropriate field. Words that did not appear in the dictionary program prompted the computer to scan the rest of the field and, if nothing had been extracted for either of the storage fields but one or more numeric words had been identified, to place the first of the numeric words in the box/street field. For example, for an address that included "Rte 3, Box 324," the computer located the "alpha" words "Rte" and "Box" in the dictionary and inserted "3" in the rural route number field, and "324" in the street/box field. If no numeric words had been found, the storage fields were left blank.

Size coding—The format program assigned a size code to each record based on the size indicators in the records when received from the sources. The code was inserted in a specific data field, depending on the source of the individual record. During record linkage, all the size codes for any record were retained by transferring the code from any record deleted as a duplicate to the appropriate field of the retained record. After linkage and deletion of duplicates, the program had the computer scan the size codes for each record; if multiple codes were present, the particular code retained depended on the size priority code for each source. These sources, and their size indicators were as follows:

Source	Size indicator
IRS forms 1040F, 1065 and 1120	Gross receipts
1982 Census of Agriculture I/S (in scope)	Total value of products sold, from 1982 census report
IRS form 941	Cash wages
IRS form 943	Payroll
1982 Census of Agriculture nonrespondents	1982 mail size code
Multiunits	All size code 15
Abnormals	All size code 16
Special lists	Varied by list (usually based on commodity inventory)
NASS list	USDA Farm Cost and Return Survey (FCRS) farm value
NASS nonfarms	All size code 17
1982 Census of Agriculture out-of-scope and 1982 Farm and Ranch Identification Survey out-of-scope	All size code 17

If no size indicator could be determined from any source, the computer assigned size code 17 ("unknown").

Possible partnership or corporation (PPC) flags—The computer program identified and "flagged" certain records as possible partnership or corporation (PPC) cases to prevent computer deletion of partnership or corporation records that had been matched with individual records. For example, John Doe might operate an individual farm as a sole proprietorship, while also having a partnership operation with Joseph Smith. In this case, the computer could match the partnership record to Doe's individual record on the basis of his name and employer identification number (EIN) and delete one or the other record as a duplicate. A PPC flag on the Doe/Smith record would change the match status to a "possible duplicate" and the clerical review would determine the final disposition.

EIN/SSN Record Linkage

General information—Employer identification numbers (EIN's) and Social Security numbers (SSN's) provided the easiest way of linking duplicate records from the various source lists. Roughly 9 out of 10 records from the different lists collected for the census mail list included an EIN, an SSN, or both. Computer programs matched these numbers on each record to other records in the files to identify (1) unique records, (2) possible duplicates, and (3) positive duplicates. When possible duplicate records were identified, they were "displayed" (i.e., printed out) for clerical review. Positive duplicates—those with matching EIN's or SSN's, matched name controls, and without possible partnership or corporation (PPC) flags—were subject to a computerized source priority code review, and the record with the higher numeric priority code was deleted from the file. (Source priority codes were assigned in reverse numerical order; i.e., a priority code of "5" meant the record had a lower priority than a record with a code of "1.")

The programs linked records based on matching EIN's to EIN's and SSN's to SSN's; records with both went through two separate linking cycles. While the EIN and SSN linking processes were carried out separately, review of possible duplicates from both was part of the general clerical review after each phase of the linkage operation in the winter of 1986-1987 and the fall of 1987.

EIN linkage—All records with an EIN were subject to the EIN linkage process. The computer program had the computer sort these records by EIN, and by PPC flag, name control, and address priority code, and then send them into the matching cycle in code priority order. That is, the record that would be deleted always entered the cycle after the record that served as the original, called the "deleting" record. The matching cycle moved the records from the sorted input file to temporary storage for the linkage operation. The computer then "wrote" the processed records to one or the other of two output files, one for records with EIN's only, and all records for deletion, and the other for records with both EIN's and SSN's (the latter would be subject to another linkage operation using the SSN's).

When the EIN's matched, the computer compared the name controls and checked for a PPC flag; if the name control matched and there was no PPC flag, the records were identified as a positive match. The sorting done prior to the linkage operation ensured that Record 2 had a lower source address priority code than Record 1, so Record 2 was flagged for deletion. The computer transferred all of the deleted record's source, size, and geographic codes to Record 1, and then read it into the appropriate output file, while a new record moved into the Record 2 location.

When EIN's matched but the name controls did not, or when one or both records contained a PPC flag, the records were declared possible duplicates. No codes were transferred, but a "possible duplicate pair" number was inserted in both records, linking them so they could be displayed together for clerical review. If Record 1 already had a pair number, the same number was inserted into Record 2; Record 1 then was written into the output file and Record 2 moved into the Record 1 location. This cycle continued until the input file was exhausted, all duplicates had been flagged, and all possible duplicates assigned pair numbers.

SSN linkage—The Bureau merged the "EIN with SSN" output file from the EIN linkage operation with the "SSN only" file to create the input file for the SSN linkage operation. The input file was sorted by SSN, PPC flag,

name control, and address priority in the same fashion as the EIN linkage input file, and the same basic linkage procedures were employed, except for the use of "dummy" file records and assignment of pair numbers.

The SSN linkage operation used "dummy" records (duplicates of the master records except that a second SSN was substituted for the original, allowing linkage of the two records) because some records drawn from the IRS 1040F file contained two Social Security numbers (usually those of spouses) and the records had to be linked to both SSN's. The computer linked only one data field for each record, hence dummy records were created for 1040F records with two SSN's. After linkage, the operation matched the dummy records to the master records for each, transferred any codes picked up during processing to the master, and deleted the dummy records.

Pair number assignment in the SSN linkage operation differed from the procedures used in the EIN operation in that there were cases in which two records were possible duplicates, but each had a different pair number assigned during EIN linkage. During the SSN linkage phase, such suspected duplicate cases retained their original pair numbers, and a secondary "collision" pair number was inserted into each record to tie suspected duplicates to the SSN-linked record.

Geographic Coding

General information—The name and address linkage operation was carried out within five-digit ZIP Code number or ZIP group number (for cities with multiple ZIP Codes), but the records in the mail file had to be geographically coded before any linkage could be done. Every record entering the name and address linkage process had standardized and edited agriculture census geographic codes, i.e., State and county numeric codes, county "alpha" (alphabetic) codes, and ZIP Codes.

Master geographic reference file—The Bureau's master geographic reference file provided the geographic codes needed to standardize and update the geographic information in the address lists. The reference file was created by combining computerized information from the ZIP Code reference file and the 1982 inscope files. The ZIP Code file listed all the post office names and ZIP Codes in the United States; each post office name entry included the standard full spelling and any known variations, as well as a fully recoded spelling, together with the State and county numeric and alpha codes, ZIP Code, and telephone area code. Matching the ZIP Code reference file to the 1982 inscope file produced a master list of unique ZIP Codes with proper and common variant spellings of most post office names, and the most likely county location for each ZIP Code—the latter based on reported primary location of the majority of farm records with that ZIP Code from the 1982 inscope file. (The computerized record for each address included two State/county geographic code fields, one for the mailed State/county geographic code, drawn from the ZIP Code reference file, and one for the reported State/county geographic code. The reported code was used to establish the census file number (CFN) for each record.) The county location was not used to code all records, since about 25 percent of the post offices listed served more than one county. ZIP Codes not matched to the 1982 inscope file retained their original county code.

Mail-file processing—Once the master geographic reference file was ready, the Bureau used it to edit the census mail file records in a series of computer operations that (1) checked the validity of the ZIP Code/post office name match on each record; (2) inserted ZIP Codes, post office names, and county and State alpha codes into records missing these items; (3) standardized spellings of post office names; and (4) assigned (mailed and reported) county and State numeric codes.

After geographic coding, the mail file was ready for name and address linkage.

Name and Address Linkage

General information—After EIN/SSN linkage and deletion, the records remaining in the mail file underwent a third matching operation using names and addresses. The name and address linkage process recoded name parts using a modified SOUNDEX system[1] similar to that used in the three previous agriculture censuses to compare names and addresses on records in the file. (The 1987 plans incorporated the 1982 improvements to the system to include the use of first and middle initials, and of numeric characters in the address.) The linkage program (1) identified name parts, (2) recoded the name in each record for linkage purposes, and (3) linked names and addresses and deleted positively identified duplicate records from the file.

Identification of name parts—The name parts in the first and second name fields in each record had to be identified before the names could be recoded. To do this, the computer compared all the words in each name field to the "skip list" (see above); words matched to words on the skip list were ignored. The computer then scanned the name fields and classified all the remaining characters and/or "character strings" (i.e., groups of two or more characters) as a surname, single letter, conjunction (e.g., "&," "and," and so on), or "other." The surname was identified using the surname locator assigned in the initial format program (see above); conjunctions were identified

[1]An indexing system that keeps together surnames of the same or similar sounds but of variant spellings. This system compensates for errors or changes in spellings over generations. The agriculture census used a modified Soundex system, that gave more weight to the specific spelling of the name, and truncated the surname for easier and more rapid access of computerized records. In this system, records bearing the names "Broom," "Bruem," and "Brume," for example, would be indexed together to check variant spellings of the name, and would be indexed under "BRM-." Other like-sounding variants, however, such as "Brougham," or "Bruham," would be indexed under "BRGH" and "BRHM," respectively.

by comparing each word to another computerized dictionary, and classifying the individual words accordingly. Each word was identified with a numeric designator (e.g., surname = "3," conjunction = "4," single letter = "2," other = "1").

After classifying each character and character string in each field, the operation retained the assigned codes, in sequence, as the name pattern. This pattern for each record then was used to identify each word or letter in the field. The computer compared the name pattern to a file of acceptable name patterns which sequentially identified each word as a first name, first initial (single letter), middle initial (single letter), or last name.

Name pattern matching rejected records primarily because the surname locator code had been set at zero, or because a particular pattern did not match one of the acceptable patterns. The latter situation occurred most frequently with multiple name strings, such as "Joseph A John B and Peter C Doe."

Name recode—With the parts of each name identified, the computer recoded the last name on each record; the first letter of each name was retained, and the second of all double characters deleted, together with all vowels (including "y"). The recoded name then was left-justified (i.e., moved to the left margin of the record) and transferred to a four-character storage cell; any excess characters, reading from left to right, were dropped from the recode, and if any recode had fewer than four characters, the last space(s) was left blank. For example:

HAMILTON became first H-M-LT-N, then H-M-LT, and finally, HMLT
SATTERFIELD became first S-T—RF—LD, then S-T—RF, then STRF
TUTTLE became first T-T-L-, then was left-justified to TTL-.

First names were recoded in the same manner (e.g., JOSEPH became J-S-PH, then JSPH; BENJAMIN, B-NJ-M-N, then BNJM; and so on), while first initials were identified and used alone. Middle names were not recoded, but middle initials were identified and used as a match key. Once the first name was identified, it was checked against a "nickname dictionary" (a list of common nicknames, such as "Bob," "Tom," "Beth," and so on); if the name was found in the dictionary, it was recoded using the proper name ("Robert," "Thomas," "Elizabeth"). Nicknames that could represent several proper names ("Ed," "Hal," "Milly") were recoded using the most frequently encountered proper name (e.g., "Edward" for "Ed"). Abbreviated names ("Geo," "Chas," "Robt") were converted and their proper names recoded.

When the computer identified a record with a multiple name pattern, it created dummy records for each possible name. Each dummy record carried all the identification codes of the original ("master") record so that it could be matched back to the master after linkage. Dummy records also were created for spouse names (except those from the IRS 1040F lists), names in the second name field, and partnership names.

For example, for a record containing in the name field "John Jones & William Smith," the recode operation identified the name pattern as "11413", which was matched to the acceptable name pattern file. The name was recoded with three possible combinations of names, "John Jones," "John Jones Smith," and "William Smith." If only single names were in the name field—e.g., "Jones, Smith, & Green"—each would be recoded with a separate dummy record to enable the linkage operation to identify partnerships that might change name order in different source file records.

Record linkage—After recoding all the master and dummy records, the computer sorted the file successively by name and address recodes within each ZIP Code group as follows: Last name, first initial, PPC flag, dummy flag, box number, rural route number, first name, and source priority code. Once sorted, the file was ready for linkage.

The name and address linkage had the same objectives as the other linking operations, to classify each record as duplicate (for deletion), possible duplicate, or nonduplicate. Six items were used to classify the records—

1. Last name
2. First initial
3. Middle initial
4. Box/street
5. Rural route
6. First name

The operation required the last name and first initial of any two records to match before making any further comparisons (records with matching last names—but no first initials or given names—were processed through the entire linkage cycle). If the last name and first initial did match, the computer compared records on each of the other key items in succession and in all combinations, and classified them based on the extent of agreement among the various matching items. The matching system classified the records, based on the following requirements:

Duplicates/computer deletes. Records matched on first and last names and on address information.

Possible duplicates. Records matched on first and last name recodes, but address information did not match or was absent. Records that matched last name and first initial, that also matched on address information, were classified as possible duplicates.

Nonduplicates. Records matched on last name recode only, or on last name recode but with different first initials. (Records matching on name and first initial recode, but with different middle initials, were classified as possible duplicates.)

When the computer identified a duplicate record during the linkage operation, it transferred the identification codes from the record with the lower source-priority code to the one with the higher priority, and flagged the low-priority record for deletion. Possible duplicates were displayed (i.e., printed out) for clerical resolution.

Clerical Resolution of Possible Duplicates

After the EIN/SSN and name and address linkage operation was completed, possible duplicates from both linkage processes were sorted by pair number and seven computer listings of possible duplicates were prepared:

1. EIN/SSN non-PPC
2. Name and address PPC cases
3. Name and address non-PPC possible duplicates
4. Combined EIN/SSN and name/address PPC cases
5. Combined EIN/SSN and name/address non-PPC
6. "Other" special sets—i.e., "collision" pairs, multiunits and abnormals, etc.

Printouts of each listing showed "sets" of possible duplicate records (two or more linked records comprised a "set") separated by lines of asterisks, each record numbered sequentially within each set with a "label position number" (LPN). Clerks reviewed the records and used written instructions to determine whether the records within each set were duplicates and designated them for deletion by circling the pair number and LPN and entering the "deleting record's LPN" (the DLPN) in the record for deletion so that the computer would transfer its identification codes to the deleting record. The clerical staff determined which was the "deleting record" by comparing the address source priority codes on each; the lower priority code record was retained. When two or more duplicate records had the same priority codes, the clerks retained the one with the most complete address information. Problem cases could be referred to analysts for resolution.

The pair number/LPN/DLPN data were keyed for all clerical deletes and the computer program processed the results of the clerical review by matching them against the possible-duplicate file.

Controls

General information—The Bureau establishes a system of checks and controls on the address list compilation operation in every census to keep track of the actual processing of the source records and to have materials available to test each phase of the operation. For the 1987 census, these controls and checks included a ZIP Code sample, a trace sample for quality control review of the overall operation, and control counts of records in the file at each processing step.

ZIP Code sample—The Bureau selected the ZIP Code sample from initial mail list input files before computer production runs began. The sample consisted of all the records in specified three- and five-digit ZIP groups within various States, and the agency planned to use it for testing each phase of the computerized formatting, linkage, and deletion processes. The plans originally were to process the samples in test runs to identify and correct any problems in the processing programs, but time constraints became so severe that only the "first cut" of the computer run (i.e., the first group of records edited by the computer in the processing cycle— between 100,000 and 200,000) could be checked.

Although Agriculture Division did not use the ZIP Code samples in the mail list preparation, as intended, the samples were also designed for potential mail list linkage research prior to the 1992 census, and were retained for that purpose.

Trace sample—The trace sample was a sample of records used to check the effect of processing on the records themselves. For the 1987 census mail list, the Bureau used the mail list compilation computer program to flag the first record and every 1,000th record thereafter in the file *prior to* input to the format and standardization operation (a total of approximately 12,000 records). When selected for the sample, and again after each step in the mail file processing operation, each record was "displayed" (i.e., printed out) and reviewed by statistical analysts. This produced a file for each sample record showing it as it entered the compilation and the changes made to it at each point in the processing. The agency's staff used the sample as a quality-control tool, and for research projects concerned with the address file processing.

Control counts—The computer used matching programs to generate control counts at each stage of the processing cycle of the number of records (1) in the input file, (2) in the output file, and (3) deleted from the file (and the stage of the cycle at which those records were deleted). The counts served as checkpoints at each phase of the mail list preparation. For example, the second phase record-linkage control counts included the following:

Count	Records
Total source records	9,174,143
EIN/SSN linkage computer deletes	2,738,335
Name and address computer linkage deletes	133,129
Clerical deletes	132,441
Other deletions	118,515
Multiunit deletions	8,566
Automatic drops (primarily nonfarm records—out-of-scope census or NASS nonfarms—that did not match another source record)	1,769,630
Model drops	174,834
Output file (i.e., final mail list)	4,098,693

Statistical Modeling

General information—The objective of the classification tree model program was to classify 1987 mail list records into groups according to their expected farm status. This was achieved by classifying 1982 census mail file records into groups based on responses to 12 questions about each record. Using 1982 Census of Agriculture information, the Agriculture Division determined the proportion of records in each group that represented farms, then applied the same procedure to the 1987 preliminary mail file records (only the dates involved changed (see below)) and created the same groups, with associated farm proportions. The division used a classification tree methodology in this program, employing information (i.e., geography, record source, and expected total value of agricultural products sold—the only common variables for all records in both the 1982 and 1987 mail lists) to predict the proportion of inscope (farm) records for specified groups from the 1987 mail list records. The modeling program resulted in dropping 175,000 addresses from the 1987 census mail file as unlikely to represent census farms.

Classification tree methodology—The classification tree methodology involved a multivariate technique to separate mail list addresses into groups according to specified classification variables, and so to predict their likely status as a census farm or nonfarm. The Bureau divided the 1982 mail list file into 29 subfiles (each including addresses for one or more States), then split each subfile in half. The classification tree procedure then was used to partition the records in the first half of each subfile into model groups according to each record's "response" to 12 questions. The queries applied to each record were "Is this record(s)—

1. A 1978 census nonfarm?
2. On a 1982 IRS list?
3. A 1978 census farm?
4. A 1978 census nonrespondent?
5. A 1978 Farm and Ranch Survey nonfarm?
6. On any 1982 special list?
7. On a 1982 USDA list?
8. 1982 expected total value of agricultural products sold (TVP) unknown?
9. 1982 expected TVP less than $2,500 or unknown?
10. 1982 expected TVP less than $5,000 or unknown?
11. 1982 expected TVP less than $60,000 or unknown?
12. A multiunit or abnormal, or has a 1982 expected TVP of $60,000 or more?"

The Bureau used the second half of each subfile to refine the classification tree based on rules for optimal classification. The refinement procedure determined which questions best divided the records by farm/nonfarm status and, in the process, obtained the minimum classification error rate. The resulting classification trees created 2,184 model groups.

The staff used information from the 1982 census to determine the proportion of addresses classified as farms in each model group. The model groups then were ranked according to *descending* expected farm proportion, numbered from 101 (numbers 0 through 100 were reserved for processing purposes) to 2,284, with model group 101 having the highest expected farm proportion.

Mail list preparation—The Bureau assumed that model groups with low farm proportion in the 1982 census would have low proportion in the 1987 census as well. The 12 questions applied to each record were modified to reflect 1987 census cycle characteristics—i.e., in all questions, references to 1978 and 1982 were changed to 1982 and 1987, respectively. The modified questions were used to place 1987 census preliminary "final" mail file records into model groups 101 to 2284. The final mail list excluded the 175,000 records in model groups with the lowest expected proportion of farms.

Source-list record linkage—The first source-list record linkage operation produced a preliminary mail list of approximately 5.9 million records, of which 1.75 million were from nonfarm sources only. These nonfarm records were retained from the second phase of linkage. The approximately 4.17 million records remaining came from the following sources:

Source	Records
Total	4,167,027
Census farms, NASS farms, IRS, and other source records	2,864,676
1982 Census of Agriculture nonrespondents and other source records	247,660
1982 Census of Agriculture nonrespondents only	272,468
1982 Census of Agriculture nonfarms and other sources	648,769
NASS nonfarms and other sources	133,454

The second linkage operation began in July 1987, matching the 5.9 million records in the preliminary mail file to approximately 3.2 million records from new source lists. After linkage, source and size codes needed for modeling and sample selection were assigned to the individual records and a file containing 6,043,157 mail list addresses was created. A total of 1,769,630 addresses from the following (largely nonfarm) sources were automatically dropped from the file:

Source	Records
Total records deleted	1,769,630
1982 census and 1982 Farm and Ranch Identification Survey nonfarms only	1,313,794
NASS nonfarms only	153,907
Census and NASS nonfarms	124,136
1982 census nonrespondents (unmatched or matched only to NASS nonfarms) with size indicator of less than $2,500 TVP	177,793

The linkage operation and automatic deletions left a preliminary "final" mail file of 4,273,527 for model application. Using the model program, the computer assigned each record to a "model group" based on source and size codes (see "Classification Tree Methodology" above), tallied the records by model group, and then split the file into three subfiles based on those tallies—"drop," "short-form eligible," and "regular census files."[2]

The Agriculture Division staff reviewed records in each file, and decided to retain 127,961 records (in the "drop" file)—those that had (1) an NASS source, (2) a 1982 census farm source (records with expected sales of less than $40,000 were eligible to receive the "short form", (the form 87A0400)); (3) a special list source, or (4) an expected total value of products (TVP) sold of $100,000 or more from any source. The remaining records in the drop file and all addresses in the "short form eligible" file that were "1982 census nonrespondents only" (or matched to NASS nonfarm only) with an expected TVP of $20,000 or less were deleted from the final mail list. Records in the "short-form eligible" file, with expected TVP's of $20,000 or more, were moved to the "regular census" list.

After these changes, a total of 174,834 records were dropped from the final mail list, leaving 4,098,693 addresses: 2,702,889 in the "regular census" file, and 1,395,804 in the "short-form eligible" file.

FINAL MAIL LIST

General Information

The final mail list preparation involved (1) assigning census file numbers (CFN's) and other processing codes to each record, (2) identifying "must" and "certainty" cases (see below), (3) selecting records to receive the sample report form, and (4) identifying cases to receive the short form.

[2]The "drop" file contained the 263,743 records in the model groups with the lowest expected proportion of farms—i.e., the calculated proportion of farms in this group was 11.7 percent or less; the "short-form eligible file" had the 1,189,265 records with the next lowest proportion of farms (11.7 to 43.22 percent); while the remaining 2.82 million or so records (addresses in groups with a minimum expected proportion of farms of 43.22 percent) were placed in the regular census file.

Census File Numbers (CFN's)

Processing the census report forms and data required a unique identification for each data record—the census file number (CFN). The Bureau assigned a CFN to each address on the final mail list. Each CFN consisted of 11 digits arranged in three groups: The first five digits were the State and county codes for the expected location of the farm, the second five a serial number identifying the specific operation within its county, and the 11th was a check digit. The check digit provided a mathematical check for quality control during data processing.

Must Cases

"Must" cases were those agricultural operations (1) so large that failure to include their data in the census tabulations would significantly distort the census data, or (2) that required special handling, such as multiunits. "Certainty" cases were expected large farm operations (based on TVP or acreage) that did not qualify as "must" cases in terms of size or type of farm, but were considered sufficiently important to justify intense followup, including telephone followup.

"Must" cases were selected by computer after record linkage was completed for the final mail list. The selection program employed size codes and lists of multiunits from the 1982 census in scope list, and other size indicators from the mail files, and identified records for (1) farms so large that some data had to be collected, rather than imputed, in cases of nonresponse; (2) operations for which an explanation was needed of why the addressee was not engaged in agricultural production; and (3) those addresses for which there were indications that the census return would need a special analyst's review. These general categories included the following types of addresses:

Multiunits. Multiunits were companies or organizations with substantial agricultural operations at more than one location. In general, a multiunit required a separate report form for each agricultural establishment, each of which was considered a separate farm for census purposes. Separate mail files were maintained for each master (i.e., company or organization) record and each associated establishment. Multiunits identified prior to the census mailout were assigned multiunit identification numbers in the alpha/plant field[3] of the address label indicating whether the report form was for the master or an associated establishment.

[3]Identified the company with a six-digit number in the alpha field of each record. The "plant" code was a four-digit establishment identifier. The master record for a multiunit would have the company identifier in the alpha field and four zeros in the plant field, while each associated establishment had the company identifier in the alpha field and a unique establishment identifier in the plant field. Each report form for a master or associated establishment was assigned a specific serial number; the associated establishments received numbers in sequence following the master.

Abnormal farms. Abnormal farms were those operated by institutions, such as State agricultural research facilities, prison farms, Indian reservations, and so on.

Other farms. The "other farms" category included addresses believed to represent large individual agricultural establishments. The size criterion (expected total value of agricultural products (TVP) sold or total acreage) used to determine "must" status varied from State to State. In Texas, for example, the minimum requirement for identification as a "must" case was a TVP of $500,000, or a total of 2,000 acres or more. In West Virginia, a "must" case required only $100,000 in sales or 1,000 acres.

Other large cases were selected for telephone followup on the basis of acreage and TVP. The minimum acreage requirement generally was the same as for the "must" category—i.e., 1,000 to 10,000 acres, depending on the specific State, while minimum TVP varied from $40,000 to $100,000. Both the "must" and "large—telephone followup" categories received intensive telephone followup during census processing. In situations in which addresses could not be contacted by telephone, or operators refused to respond, secondary sources, such as the USDA's Extension Service (ES) and/or Agricultural Stabilization and Conservation Service (ASCS) offices were asked to provide information as to whether nonrespondent addresses had agricultural operations. Data from previous census records, in conjunction with other information, were used to impute responses for nonrespondent addresses.

Mail List Sampling

The Census Bureau introduced sampling for data collection in the 1945 agriculture census, but did not use it again, except for selected post census and research surveys, until the 1978 enumeration. The 1978, 1982, and 1987 censuses sampled to collect specified additional data from selected agricultural operations; all farms were asked for basic data, with an approximate 25-percent sample of the mail list sent a "sample" form that requested additional information on such items as production expenses, use of fertilizers and insecticides, value of machinery and equipment, and so on. To further reduce overall response burden, in the 1987 census the Bureau also employed a "short" form (one sheet, front and back) with abbreviated versions of the standard data items. Addresses less likely to meet the census farm definition received the short form.

The sampling technique used in the 1987 census was virtually identical to that of 1982: During mail list compilation, addresses were classified as "certainty" or "noncertainty" based on expected value of sales of agricultural products and acreage (these varied by State), and including all multiunits and abnormal operations. After linkage and unduplication, and statistical modeling of the final mail list, the "regular census" and "short-form eligible" files were merged and sorted by CFN for sample selection. The sample included all certainty addresses, all addresses in Alaska and Hawaii, all addresses in counties with fewer than 100 farms in the 1982 census, and a stratified sample, by county, from the remainder of the mail list. The sampling rate for each county was determined by the total number of farms in the county in the 1982 agriculture census—counties with 100 to 199 farms were sampled at a 1-in-2 rate, and counties with 200 or more farms at a 1-in-6 rate.

After sample selection, the "short-form eligible" file (excluding cases selected for the sample) was sorted by model group according to descending farm proportion, and the first 906,406 records were selected in sequence to receive the short form. The final mail file was as follows:

Report form type	Records
Total	4,098,693
Sample/certainty	1,107,452
Nonsample	2,084,835
Short	906,406

PRINTING AND ADDRESSING REPORT FORMS

General Information

Private contractors printed the report forms and various mailing materials and prepared the mailing packages for the 1987 agriculture census.[4] The contractors printed all the materials and assembled the mailing packages to agency specifications, under quality-control supervision by Census Bureau personnel, and then forwarded the packages to the Jeffersonville, IN, facility for final preparation (primarily address labeling) and mailout.

Address Labels

The Bureau prepared the address labels for the 1987 agriculture census mailout "in house." The census mail address list was generated at the main computer facility in Suitland, MD, then transmitted to Jeffersonville, IN, by telephone datalink and copied onto computer tape. The Data Preparation Division (DPD) staff in Jeffersonville used the address list tapes to print the census address labels on high-speed printers.

Printing, Assembling, and Addressing

General information—The seven private contractors employed in printing most of the report forms and other mailout materials also assembled the mailing packages according to Census Bureau specifications. Each contractor printed all of the materials for a specified package (e.g., sample,

[4]Report forms and mailing package materials were printed by seven contractors while eight produced the mailing envelopes and supplied them to the report form contractors for packing. Most of the major contractors were located within 300 miles of the Jeffersonville, IN, office; this facilitated easy delivery of the mailing packages to the Data Preparation Division for labeling and mailout.

certainty, nonsample, or "must" case for a specific geographic region), assembled the packages in envelopes supplied from the envelope contractor, and delivered them to the Jeffersonville office. The contractors supplied complete mailing packages for (1) the initial mailout, (2) the followup mailings, (3) additional sample, nonsample, and "must" packages for mailing to postmaster return (PMR) cases and "adds," and (4) all types of report forms as general reference materials and for mailing to correspondents or respondents on request.

The staff at the Jeffersonville office inspected the mailing packages as part of the quality control program, added any special instructions needed for specific packages (e.g., for such operations as bee and honey producers, contract poultry operations, and so on), and applied address labels for the mailout.

Quantities—The total number of standard report forms printed for the 1987 agriculture census was as follows:

Region	A0400 (Short)	A01 (Nonsample)	A02 (Sample)	A03 (Must)
Total	2,250,000*	6,333,000	2,398,500	620,500
01		399,000	207,000	56,000
02		1,327,000	416,000	100,000
03		577,000	216,000	57,000
04		1,284,000	456,000	90,000
05		134,000	52,000	18,500
06		896,000	324,000	76,000
07		533,000	158,000	56,000
08		278,000	137,000	43,000
09		178,000	83,000	33,000
10		267,000	85,000	41,000
11		310,000	121,000	50,000
12		-	33,000	-
13		-	10,500	-
14		150,000	100,000	100,000

*The 87-A0400 was a standardized form; no regionalized versions were produced.

A facsimile of a representative report form is included in appendix F.

Other printed materials ordered for the data collection mailings included the following information sheets, form letters, and envelopes:

Form number	Description	Quantity
Information Sheets and Form Letters		
87-A01(I)	Information sheet	9,391,000
87-A02(I)	Information sheet (Hawaii)	29,000
87-A04(I)	Information sheet (for A0400 "short" form)	2,250,000
87-A01(L1)	Transmittal letter (initial mailout)	4,939,000
87-A01(L1A)	Transmittal letter (PMR's)	140,000
87-A01(L2)	Reminder card	8,800,000
87-A01(L3)	Followup letter	3,079,000
87-A01(L4)	Followup letter	1,950,000
87-A01(L5)	Followup letter	1,652,500
87-A01(L6)	Followup letter	1,300,000
87-A01(L7)	Followup letter	1,186,000
Special Instruction Sheets		
87-A31(A)	Grazing associations	1,100
87-A31(B)	Institutional organizations	5,000
87-A31(C)	Indian reservations	350
87-A31(D)	Farm with multiple farm and ranch operations	6,000
87-A31(E)	Contract poultry producers	45,000
87-A31(F)	Bee and honey production	150
87-A31(G)	Feedlot operations	13,000
87-A31(H)	Fish and other aquaculture	3,300
87-A31(I)	Laboratory animal producers	450
87-A31(J)	Nursery and greenhouse crops	63,000
87-A31(L)	Citrus caretakers	9,000
Envelopes		
87-A7A	Outgoing envelope (initial mailout)	5,000,000
87-A7B	Outgoing envelope (followup)	6,560,000
87-A7C	Outgoing envelope (general)	700,000
87-A7D	Outgoing envelope (initial mailing, Hawaii, Alaska, and multiunits)	30,000
87-A7E	Outgoing envelope (followup, Hawaii, Alaska, and multiunits)	60,000
87-A8	Return envelope	12,350,000

Facsimiles of the general information sheet, transmittal letter, reminder card, and principal followup letters are included in appendix G.

The contents of the initial mailing packages for nonsample, sample, and "must" cases were as follows:

Type	Report form	Information sheet	Return envelope	Cover letter
Nonsample	87-A0101 through -A0111*	87-A01(I)	87-A8	87-A01(L1)
Sample	87-A0201 through -A0213*	87-A01(I) or 87-87-A02(I)	87-A8	87-A01(L1)
Must	87-A0301 through -A0311*	87-A01(I)	87-A8	87-A01(L1)

*As appropriate.

Quality control—Teams of two or three DPD quality control (QC) personnel inspected daily each of the private contractors' printed materials and assembled packages. Report forms and envelopes were subject to a "press inspection" (a visual review to make certain the printing was of acceptable quality, the proper colors and shading were used, and so on), while the QC staff checked a random sample of assembled mailing packages to ensure that the packages were complete and the materials inserted in the right order.

Each contractor boxed each day's production of assembled mailing packages for QC review. The QC staff inspected packages from each day's production, picking three packages at random from each box selected. If the production lot consisted of 9 or fewer boxes of packages, all the boxes were sampled; for lots of 10 to 150 boxes, 5 boxes—selected at random— were sampled, while for lots of 151 to 1,200 boxes, 20 were sampled. In any lots of over 1,200 boxes, 32 were sampled. When the staff identified an error in the packaging, the rest of the packages in the affected box were inspected and, if the problem was found in other packages, the surrounding boxes were checked as well.

The most frequently encountered problems in the package assembly operation were (1) insertion of the mail package contents in the wrong order, (2) failure to seal the outgoing envelope, and (3) use of the wrong report form for a specific package.

All detected errors were corrected before the packages were shipped to Jeffersonville for labeling and mailout.

Multiunits and abnormals—During the agriculture census mail list compilation operations, the Agriculture Division identified multiunits (i.e., companies or other organizations with two or more independent farm operations) and abnormals (institutional farms) and established a separate mail file for them. In November 1987, the multiunit and abnormal computerized address file was transmitted to Jeffersonville and labels were printed. The agriculture census unit in the Jeffersonville office manually assembled and labeled the mailing packages. Treating multiunits and abnormal farms as "must" cases, it assembled the mailing packages with the appropriate regional "must" report forms for approximately 9,000 multiunit and abnormal addresses.

Enumeration packages for abnormal cases were mailed and followed up as part of the general census mailing; packages for multiunit cases were part of the initial mailout, but were followed up separately.

Labeling—Labels for the mailing packages were printed by form number in ZIP Code sequence. Four labeling machines at the Jeffersonville, IN, office addressed the packages by applying the adhesive labels through the open window on the front of the form 87-A7A outgoing envelope. The machines applied the labels at the rate of up to 10,000 per hour between the last week of October and the first week of December 1987. The Bureau released all of the over 4 million mailing packages for the census mailout to the Postal Service between December 16 and December 21, 1987.