1992 CENSUS OF AGRICULTURE FRAME DEVELOPMENT AND RECORD LINKAGE

Tommy W. Gaulden, Jane D. Sandusky, Elizabeth Ann Vacca, U.S. Bureau of the Census
Tommy W. Gaulden, U.S. Bureau of the Census, Washington, D.C. 20233

KEY WORDS: Duplicate, farm, list, completeness

1. INTRODUCTION

The 1992 Census of Agriculture, currently being conducted, is the sixth quinquennial agriculture census for which data have been collected through a mail enumeration. Due to the difficulties and expense involved with maintaining a current list, a new list frame is constructed for each census by procuring and linking files from multiple sources.

The population of farms to be covered in an agriculture census is highly diversified with respect to many characteristics, including acreage, value of production, and type of organization. Most farms are sole proprietorships, but institutional farms and farms operated by corporations and partnerships under a wide variety of arrangements also have substantial economic importance. Under the current farm definition, in use since 1974, a place with as little as $1,000 in sales (or potential sales) of agricultural products during the reference year qualifies as a farm. In the 1987 census, farms with sales of less than $2,500 represented 23.5% of all farms but accounted for only 0.4 percent of the total value of agricultural products. Farms also vary widely in their types of production, ranging from general crop and livestock farms to highly specialized units such as nurseries, feedlots, citrus groves, and dairies.

Since the completeness of the census depends on how well the mail list covers this population, it follows that the development of the census mail list is a critical step in the overall process of taking the census of agriculture. The mail list for the 1992 Census of Agriculture was compiled from several large statistical and administrative record source lists, along with several smaller special lists. These source lists generally contain names of individuals, businesses, and organizations which are associated with agriculture; many records are included which do not actually represent farm operations eligible for census enumeration.

The goal of the mail list development process is to compile a complete list - one covering the population of farm operations - while, at the same time, minimizing the number of duplicate records and nonfarm records. This goal is complicated by several factors:

(1) Some farm operations, especially the marginal ones, will not appear on any of the source lists for one reason or another.

(2) Many farm operations appear on multiple sources. The difficulty in accurately identifying duplicate records sometimes results in farm records being removed from the mail list erroneously as duplicates. Conversely, duplicate records which go undetected may be left on the list.

(3) The difficulty in accurately identifying nonfarm records sometimes results in farm records being removed from the mail list as nonfarms, or nonfarm records being left on the list erroneously. The existence of many small marginal operations contributes to this problem. An operation might qualify for enumeration one year but not the next.

The 1992 census mail list development process took place in two phases. The first phase, using source files which were available by late 1991 produced a preliminary mail file. The second phase used additional source records not previously available and resulted in the final list. In each phase, duplicate records were detected by linking records by social security number (SSN), employer identification number (EIN), and name and address information. After the second phase, a classification model was used to help reduce the file by eliminating likely ineligible (nonfarm) records.

2. PREPARATORY PROCESSING

Efforts are made to include all important sources of agricultural information in the list compilation. This includes records from the Internal Revenue Service (tax records of employers, businesses, and individuals with farm

income), the U.S. Department of Agriculture (statistical records of the National Agricultural Statistics Service, NASS, and the Agricultural Stabilization and Conservation Service, ASCS), previous census records, and private records of agricultural trade associations and other organizations. For both phases, a total of almost 12.5 million source records went into the process.

In addition to these lists of probable farm operators, lists of identified nonfarms were used to aid in the identification and removal of nonfarm records by matching to other records. These include nonfarms identified from the previous census and from NASS surveys.

Special procedures were used to place each source record into a standard format and for generating processing code fields (Dea, et al, 1984). These computerized procedures include a source record edit, assignment of name control, assignment of processing codes/flags, and size coding.

The source record edit removed any commas, periods, and most special characters. It also inserted a space between any adjacent alphabetic and numeric characters.

Name control is defined as the first four characters of a key word in the name field (usually surname). It is used in determining duplicate status during the identification number linkage, as well as in the name parsing routines prior to name and address record linkage. Special tools were developed to aid in establishing an appropriate name control. One such tool was a "skip list" dictionary. This dictionary contained over 1,000 words and abbreviations (such as FARM, DAIRY, BROS, etc.) which could conceivably appear in the name field but were not likely to be the surname.

Processing codes were assigned in the initial record format and standardization to facilitate the use of the most reliable information in the final record. In particular, each record was assigned a name and address priority code, which was used in the linkage process to determine which source record to retain in the case of duplicates. The priority code was based on the expected currency of the record source address information.

The record linkage process was designed to prevent computer deletion of partnership or corporate records matched with individual records. Since individuals are commonly involved in both partnership and sole proprietorship operations, source records that possibly represented a partnership or corporation were identified and assigned a "PPC" (possible partnership or corporation) flag. This flag would be used to prevent erroneous computer deletion of separate operations, permitting a clerical decision to be made on the linked records. PPC flags were assigned based on the source of the record and the presence of certain words associated with partnerships and corporations identified on the "skip list" word dictionary.

Each record was also assigned a measure of estimated size derived from size indicators present in the source record. The size code is expected to be an estimate of the total value of agricultural products (TVP) that were sold or could be sold in the census year. Each source had a separate field for this size code, so that during record linkage the size code would be retained for all sources on which a name appeared by transferring data from the deleted duplicate record to the retained record. The final "source combination" and "mail size" codes derived from these are important variables used in the classification model, in census processing, and in evaluating the census mail list.

A geographic coding system was designed to ensure that all records entering the record linkage system contained standardized and edited geographic codes.

3. RECORD LINKAGE

EIN and SSN Record Linkage: The most effective means for linking records from the various sources was to match on the EIN and the SSN. Over 7.8 million of the 9.1 million records formatted for the first phase (85.5%) contained at least one of these identification numbers. Whenever records matched on the EIN or SSN, a further comparison of the records was made and one of two outcomes resulted:

(1) The records were classified as DUPLICATES if their name controls were equal and neither record contained a PPC flag. When this outcome occurred, one of the records (as determined by the address priority codes) was flagged for deletion. All of the source-size codes contained in the record to be deleted, along with various other data, were

transferred to the deleting record, provided the corresponding fields were blank in the deleting record.

(2) The records were classified as POSSIBLE DUPLICATES if their name controls were not equal or if a PPC flag was present in either record. When this outcome occurred, a possible duplicate pair number was assigned to the records so that they could be displayed as a set for clerical resolution.

Name and Address Record Linkage: A record linkage program based on the names, address information, and other identifying information was used to further reduce duplication in the file. It was desirable to delete as many true duplicates as possible, yet retain records which represent separate agricultural operations. Like the EIN and SSN matchers, this linkage program implemented a decision rule designating some records as duplicates and some as possible duplicates. Unlike the EIN and SSN linkage, the decision rule is flexible; that is, since it is parameter driven, the outcomes can be changed by adjusting parameters.

In preparation for this linkage program, the mail file was processed through the geographic coding operation and a program which parsed the name and address fields to identify name parts and other variables to be used in matching. Then the record linkage was performed separately within a ZIP code or group of ZIP codes. The name and address parser used a special purpose routine which identified name parts through a word dictionary and a name pattern coding scheme, based on the type of words and their sequence (Dea, et al, 1984).

Record linkage occurred separately within each "block". A block consisted of all records within a ZIP code (or ZIP code group) with the same first character of surname. The ZIP code group was used to combine all records for a multi-ZIP city into the same block. With the exception of repeated pairs and records previously flagged for deletion, all pairwise comparisons of records within each block were considered. Each record pair was assigned a match weight which was computed from the extent of agreement between their respective match variables - surname, first name, middle initial, box/house number, rural route number, street name, phone number, and SSN.

(1) Basis for the Decision Rule:

Let **v** denote the agreement-disagreement pattern resulting from the comparison of a particular pair of records. **v** is a binary vector of dimension 8 (the number of match variables). A value is 1 if the two records agree on a match variable and 0 otherwise (Thibaudeau, 1992).

Let m(**v**) be the probability of observing **v** given that the pair of records generating **v** has a true match status, and let u(**v**) be the probability of observing **v** given that the pair generating **v** has a true non-match status. It has been shown that the pairs most likely to represent true matches are those generating the vectors **v** which maximize the ratio R = m(**v**) / u(**v**) (Winkler, 1990).

(2) Match Weight Assignment:

The weight computation consisted of two components. The initial weights were defined as the natural logarithms of the likelihood ratio, R. In practice, we do not know the probabilities associated with R. These were estimated using an Expectation-Maximization (EM) algorithm. Starting with an initial estimate of the probabilities based on previous work, and assuming that the probabilities of agreement are independent among the match variables, the EM algorithm refines the estimates of R.

The second component consisted of adjusting the initial weight based on expert judgment. This adjustment was needed because the EM algorithm is less accurate for finding probabilities associated with rare events such as agreement on SSN. Weights could be increased when this occurred. Also, weights could be adjusted downward; for example, when one of the records in a pair contained a PPC flag.

(3) Application of the Decision Rule:

The final weight determined the classification of each record pair as duplicates, possible duplicates, or non-duplicates. These were determined by two parameters (a high and a low cutoff) which were set for each group of records processed through the matcher. Pairs having weights above the high cutoff were designated as duplicates, while pairs with weights below the lower cutoff were non-duplicates. Pairs with final weights falling in between were designated as possible duplicates for clerical review. This

flexibility allowed us to set the cutoff values so as to maintain an acceptable level for false matches, while preventing an excessive workload for the clerical review operation.

An additional feature of the name and address record linkage process was the use of string comparators. The string comparators take into account a partial agreement between some match variables by considering the agreement on a character-by-character basis.

Clerical Review: Sets of possible duplicate records were initially formed during the EIN and SSN record linkage programs. Existing sets could be enlarged during the name and address record linkage. They could also be eliminated (in the case of duplicate classification), or new sets could be formed. Most possible duplicate sets were reviewed and processed interactively by clerical personnel using large-screen terminals. The reviewers, using specific guidelines, identified which, if any, records in a set to delete. A small proportion of the possible duplicate sets were selected for resolution through telephone contact.

4. CLASSIFICATION ANALYSIS

Due to budget constraints and efforts to reduce respondent burden, limits were placed on the size and composition of the final census mail list. The total mailout was limited to 3.55 million records, of which no more than 3.2 million could receive either of the regular census report forms (8-page short form or 12-page long form). Up to 600,000 records could receive a screener form (short form with initial screener section to allow nonfarm operations to skip out and return the form). After both phases of record linkage, the mail list contained approximately 5.0 million records. Deletion of 1.2 million "automatic deletes" (unmatched nonfarm source records) brought the list down to 3.8 million addresses. The final stages of processing used a classification analysis to further reduce the mail list to the required 3.55 million records.

Classification analysis is a nonparametric method of classifying records and was performed using Classification and Regression Tree (CART) software (see California Statistical Software, Inc., 1985). The basic method involved using 1987 Census of Agriculture mail list records and data to develop prediction rules/models. The prediction models were applied to the 1992 mail list records so that those records which were least likely to represent farms could be excluded from the mail list or designated to receive the census screener report form.

The classification models were developed using only record characteristics common to both the 1987 and 1992 list records. This limited the variables to geographic location, the estimated size, and the source(s) of the record. The geographic location was used to perform the analysis at the state level. Estimated size is based on information contained in the source records and is an indicator of the expected total value of agricultural products sold (TVP) by each farm. There were 14 possible record sources and 17 size codes for classifying the records. A descriptive vector was created which uniquely defined the source and size characteristics of a particular record. Based on the presence or absence of each variable on a record, the 1987 mail list records were separated into "model groups". Each model group had an associated descriptive vector and an associated farm probability. The farm probability is the proportion of 1987 farms in the model group. (see Owens, et al, 1989)

With the model groups defined, including the descriptive vector, the size and source for each 1992 mail list record was matched to the descriptive vectors. Once a match was found, the 1992 record was assigned the associated model group and farm probability. The 1992 classification analysis created 787 model groups of the approximately 3.8 million records. Those model groups containing the lowest farm probabilities were targeted by the classification analysis to be excluded from the mail list to reach the 3.55 million record limit. This resulted in the following preliminary files:

Cases to be Retained on the mail list..... 3,564,220
Cases to be Dropped from the mail list.. 219,082

State tables by model groups, mail size and selected source combinations were produced for review by Agriculture Division personnel. After review of these tables a consensus was reached for subjectively shifting specified groups of records between the files. These shifts were as follows:

From mail list file to drop file..... 145,026
From drop file to mail list file..... 134,445

Once the final list was defined, the model was also used to designate records to receive the census screener report form. These were selected from the records with estimated TVP less than $25,000 and having the lowest farm probability.

5. RECENT IMPROVEMENTS AND SUMMARY

The basic agriculture list development methodology used for 1992 was first developed for the 1978 Census of Agriculture. With each successive census cycle have come efforts to refine and improve various aspects of the process. For example, the use of nonfarm records from the previous census was introduced in 1982 as a means to identify likely nonfarm records from other sources through matching. Prior to the 1987 census cycle, a large screening survey was used to identify and remove nonfarm records from the list. The necessity to eliminate the costly screening survey led to the development and implementation of the classification model for the 1987 processing cycle. Improvements for 1992 were aimed at streamlining the manual review operations, improving duplicate detection capabilities, and refining the classification model.

Automated Clerical Review: Traditionally, the clerical review operation involved the printing of thousands of pages of computer printout. After clerical processing of the possible duplicate sets, the listings were sent to a keying operation where the clerical actions (entered on the listings) were captured for computer processing.

For the 1992 processing, an interactive computer system was developed for this operation. Possible duplicate sets were displayed for review by the clerks. Once a decision was made, the clerk identified any records to be deleted by interactively keying an action code. This method not only eliminated the need for printing and control of the paper listings, it also eliminated the separate data entry operation. This method proved highly successful in terms of cost and timeliness. The following data compares the 1987 and 1992 manual review processes:

Item	1987	1992
Sets Processed.....	1,100,900	769,267
Records Processed.....	2,430,019	1,979,936
No. of Person Hours.....	13,960	6,531
Total Cost ($1,000).....	227	106
Cost Per Set (Dollars)...	0.21	0.14

Duplication Telephone Survey: The clerically reviewed sets involving potential partnership or corporate type records are very difficult to resolve with a high degree of accuracy. For example, the same individual may be involved in two separate operations - their own individual operation and as a partner in another. By design, we require that both records be kept on the list, although this contributes to duplication in cases where only one operation exists.

Another innovation for the 1992 process was to resolve some of these sets by telephoning the individual(s) involved. To remain within budgetary constraints, we selected a sample of 25,000 PPC sets for resolution in a telephone operation. Although additional evaluation is pending, the results were encouraging. Over a period of five weeks, a group of seventeen interviewers, working only during the day, were successful in resolving a high proportion of the selected sets.

Record Linkage Methodology: Major enhancements were made to the name and address record linkage methodology for 1992. Ever-increasing budgetary constraints lowered the limits on the mail list size and necessitated greater emphasis on efforts to increase duplicate detection. The name and address matcher used in the past employed decision rules based on empirical studies of matched records. Blocking was based on recoded surname within a ZIP code or ZIP group. A limited number of variables - surname, first name, middle initial, box/house number, rural route number - were used as match keys. This matcher was highly effective in detecting the more obvious duplicates, and was accurate in terms of a low false match rate. However, it did not detect a high proportion of the total duplicates and forced too many records into the "possible duplicate" category.

The name and address matcher used for 1992 was based on a probabilistic linkage model,

developed by the Census Bureau's Statistical Research Division, and modified to fit the needs of the agriculture list application. This matcher, by using weights produced by the EM algorithm along with weight adjustment based on expert judgment, is more discriminatory in the search for duplicates. The new matcher also extracted and compared more information from the records under scrutiny (e.g. street name, telephone number, and SSN). Blocking was based on the ZIP code and first character of surname, thus providing for more comparisons. Also, string comparators were used to compare the names, telephone numbers, and identification numbers. The string comparators can discern the degree of similarity between two strings of letters or numbers. Using the match weights produced by the EM algorithm allowed us to reduce the clerical review workload by adjusting the upper and lower cutoff values.

CART Software: The classification model developed for the 1987 census list used software which was developed and programmed in-house. Although highly effective, we believed that many efficiencies could be gained from a customized software package. The software package, "Classification And Regression Trees" (CART) developed by California Statistical Software, Inc. was used for the 1992 classification model. This software package provided for more efficiency and flexibility, at a lower cost, than our in-house programs.

Summary: Building an accurate and comprehensive frame for a data collection operation as large as the census of agriculture is a difficult task. The methodology used in this process is effective to the extent that it successfully identifies and removes most duplicate operations and nonfarm records from the list. But census coverage will suffer if qualifying operations are eliminated erroneously at the mail list development stage. The frame development process must balance these concerns, as well as have a relatively low cost due to the large number of records involved.

It is unlikely that a census of agricultural operations enumerated exclusively from a list frame will ever include all operations meeting the current farm definition. Coverage of all farms has generally been around the 90 percent level. However, the list coverage for larger farms has usually been above 95 percent and for the value of agricultural economic activity about 98 percent. To maintain even this level of list coverage will require continual attention to future improvements.

6. ACKNOWLEDGEMENTS

The authors wish to acknowledge the assistance of a number of colleagues for their help in preparing this paper. In particular, we wish to thank Nash Monsour, Assistant Division Chief for Research and Methodology in the Agriculture Division for his support, and Paul Lewis and Anthony Williams, Census Bureau reviewers, for their helpful comments.

REFERENCES

1987 Census of Agriculture, U.S. Summary and State Data, Volume 1, Part 51.

1987 Census of Agriculture, Coverage Evaluation Report, Volume 2, Part 2.

California Statistical Software, Inc. (1985), "An Introduction to CART Methodology".

Dea, Jane Y., Tommy W. Gaulden and D. Dean Prochaska (1984), "Record Linkage for the 1982 Census of Agriculture Mail List Development Using Multiple Sources," 1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.

Owens, Dedrick, Ruth Ann Killion, Magdalena Ramos and Richard Schmehl (1989), "Classification Tree Methodology for Census Mail List Development," 1989 Proceedings of the Section on Survey Research Methods, American Statistical Association.

Thibaudeau, Yves (1992), "Identifying Discriminatory Models in Record Linkage," 1992 Proceedings of the Section on Survey Research Methods, American Statistical Association.

Winkler, William E. (1990), "Understanding Record Linkage," U.S. Census Bureau internal report.