Removing Duplication from the 2002 Census of Agriculture

Kara Daniel, Tom Pordugal

United States Department of Agriculture, National Agricultural Statistics Service
1400 Independence Ave, SW, Washington, D.C. 20250
kdaniel@nass.usda.gov, tpordugal@nass.usda.gov

Key Words: Record Linkage, List Sampling Frame, Duplication

Introduction

In 1997 the responsibility for conducting the Census of Agriculture was transferred from the Bureau of the Census to the National Agricultural Statistics Service (NASS). With that transfer, NASS became responsible for providing County, State and U.S. level estimates of agricultural production, value, costs, and demographic characteristics.

Data for the Census of Agriculture are collected by mailing questionnaires to all known and potential farms in the United States. The complex characteristics of farm operations make compilation of a complete and efficient Mail List for the Census of Agriculture a challenging process. The turnover rate, particularly for small farm operations, is high. Based on longitudinal comparisons from one Census to the next, small farms exit at an annual rate of 15%. Over the same period, the exit rate for large farms is approximately 2-3%. Another factor that complicates the build process is the fact that agricultural operations are commonly associated with multiple individuals and/or addresses. Furthermore, agricultural operations are often known by several different operating names.

The Census Mail List for the 2002 Census of Agriculture was generated using the NASS List Frame. Time and effort were devoted to producing a list with the coverage of the farm population and lowest duplication level possible. However, it is known that achieving a list with 100% coverage and no duplication is not possible given time and resource constraints. Analysis of the 1997 Census of Agriculture found coverage levels at approximately 87%. Duplication levels were not published. Coverage and duplication levels for the 2002 Census of Agriculture are not yet available. High coverage and low duplication are important for accurate estimates. Low duplication is also important to minimize respondent burden.

This paper will focus on the methodology used by NASS to identify and remove duplication as part of the 2002 Census of Agriculture. It will detail efforts made to remove duplication before the Census Mailout and after Census data were received. It also will describe the record linkage methodology for matching outside source data to the NASS List Frame and for unduplicating the NASS List Frame. Finally, it will present some recommendations of areas where improvement can be made in the future to reduce the duplication levels further.

Duplication in the 2002 Census of Agriculture

Building the Mail List

The transfer of responsibility for the Census of Agriculture from the Bureau of the Census to NASS brought changes in the way the Census Mail List was developed. In prior Censuses, the Bureau of the Census compiled the Mail List by amassing information from a variety of agricultural list sources. The NASS List Frame was used as one of the input sources. After compiling all the different lists and transforming them into a standard layout, the sources were then unduplicated using automated record linkage. The process of building the Mail List was done just prior to the Census Mailout. Following the completion of the 1997 Census, the 1997 Census Mail List and accompanying Census Data were merged with the NASS List Frame. The single list, maintained by NASS, was the frame for the 2002 Census of Agriculture and the sampling frame for NASS's ongoing survey program. The Mail List for the 2002 Census of Agriculture was obtained by pulling all known farm and potential farm records from the NASS List Frame.

Rather than a single large scale list building effort, NASS builds and maintains its List Frame continually using survey data as well as data from a variety of administrative lists. Names, addresses, phone numbers and agricultural data items are constantly updated as more current information is obtained. It is hoped that NASS's list maintenance work will result in a frame with maximum coverage and minimal duplication. As outside source lists are obtained, records are checked to see if they overlap with the NASS List Frame. Records found on the frame are updated if more current information is available. Records not found on the frame are added with a status code indicating that the operation is a potential farm operation. Depending on the size and medium of the lists, they are either reviewed manually or using probabilistic record linkage. The probabilistic record linkage process checks both for overlap between the outside source records and the NASS List Frame and for duplication within the outside source records.

Following the 1997 Census, analysis of the NASS List Frame was done to determine areas where List Frame coverage was the weakest. The analysis showed that coverage of CRP only farms (operations whose only activity is participation in the Farm Service Agency's Conservation Reserve Program), specialty livestock farms, and equine farms were weakest. Special effort was put into obtaining outside source lists targeting these weaker areas at both the Headquarters and State levels. In preparation for the 2002 Census, outside source lists were obtained from a variety of sources. These sources included State Farm Census lists, federal lists, breeding association lists, livestock or crop association lists, farm bureau lists, seed grower lists, veterinary lists, marketing association lists, and a variety of other agricultural related list sources. Lists from almost 50 different outside sources were run through the record linkage process. Many lists had data for more than one State. These lists resulted in almost one million new potential farms being added to the NASS List Frame. The status of 320,000 existing non-farm records was updated to indicate that the records were potential farms.

NASS Record Linkage Procedures

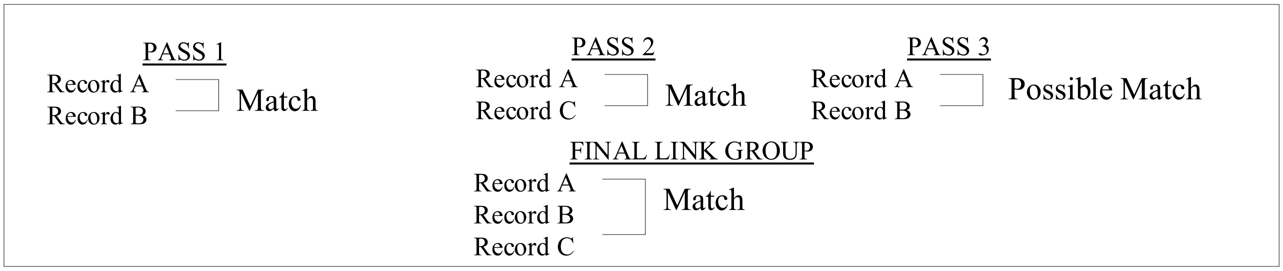
To build its frame and remove duplication, NASS conducts many record linkage projects every year. These projects are primarily used to match outside source lists to its List Frame and to unduplicate outside source lists and to unduplicate the NASS List Frame. With each record linkage project, the NASS List Frame is treated as the reference frame. Records are added, updated or dropped from the NASS frame based on the results of each match. Records are linked together using probabilistic record linkage methodologies executed by Headquarters staff. NASS currently is using Ascential Software's SUPERSTAN and SUPERMATCH software as its engine for standardizing and matching lists. The Ascential Software was built based on theory proposed by Fellegi and Sunter in their 1969 JASA article. [1]

NASS's record linkage projects are all run on a State-by-State basis. Records from outside source lists are both matched to the current NASS List Frame and to themselves (unduplicated). Both the match and unduplication are done through a series of passes using different blocking and matching variables. Each pass identifies a set of records linked as matches, possible matches, and nonmatches.

Certain variables, like Social Security Number (SSN), Employer Identification Number (EIN), and phone number, are considered key matching variables. It is important that records with the same values for key variables are brought together. To do this, independent passes are run for each key variable with that variable used as a blocking variable. Supporting variables such as surname or house number are used to separate the matches and possible matches. A simplified example of a key variable pass would be a pass blocking on SSN and matching on surname. All records with the same SSN would either be classified as matches or possible matches. Pairs of records with the same SSN and surname would be considered matches. Pairs of records with the same SSN and different surnames would be considered possible matches. NASS also includes a series of other more typical probabilistic record linkage passes where records are blocked and matched on a wide range of variables.

In a traditional record linkage project, the possible matches from one pass are reviewed before running subsequent passes. However, NASS has modified the typical processing methodology to better meet its needs and to reduce overall review time. The methodology currently used by NASS includes running each pass with all records from both files being input into each pass. The output files from all passes are then combined using a SAS program. This program combines different sets of linked records with one or more common records into one big link group. If two records are linked as possible matches in one pass and the same two records are linked as matches in another pass, they are classified as matches. If any subset of records in a link group were classified as possible matches, the entire link group would be considered a possible match group.

The following image displays an example of how output from three different passes would be combined into a final link group by the SAS program. In the example, records A and B were linked as matches in pass 1. Records A and C were linked as matches in pass 2 and records A and B are linked as possible matches in pass 3. Passes 1 and 3 resulted in the same two records being linked to each other. However, the records were matches in pass 1 and possible matches in pass 3. Though records A and B were classified as possible matches in pass 3, they would be considered matches by the SAS program because they had enough common information to be considered matches in pass 1. The links generated by all three passes included record A. Because of this commonality, the SAS program would combine the records from each pass into one large link group containing records A, B, and C. The final link group would be considered a match because the link between records A and B was a match and the links between records A and C was a match.



There are two primary reasons why NASS has deviated from the traditional methodology of reviewing record linkage output on a pass by pass basis. The first reason is that management of the review process is much easier with one combined review as opposed to a set of multiple reviews after each pass. Personnel in NASS's field offices usually review the link groups for their State(s). If a given project included 9 passes, there would be 414 different passes (9 passes * 46 field offices) which would have to be monitored. As a State completed the review of one pass, the next pass would have to be run and then reviewed. Running all passes at once and having the field offices complete one review is more efficient than multiple runs and reviews. One disadvantage of combining the records from all passes is that there is potential to have a large number of records in a link group. The second reason NASS has opted for one combined review is that NASS has learned that the links identified in one pass are often helpful in resolving the links from another pass. For example, records A and B may have been identified as links based on the same EIN in one pass and records A and C may have been identified as links based on the same SSN in a different pass. Since all three records potentially contribute information to one farm operation, it is helpful to see them together in the review process. With a combined review, records related to the same operation only have to be reviewed once rather than multiple times across passes. A combined review also provides a more complete picture of the situation. This is particularly helpful when follow-up calls are needed.

Linked records are reviewed using NASS's in-house resolution system. Field Office staff typically review all possible matches and a portion of the matches and nonmatches. Having the review performed in the State Offices is advantageous because of the vast knowledge State personnel have with the farm operations in their particular State. State personnel have resources to attempt to contact operations when questions arise as to the status of an operation. NASS's clerical review system allows reviewers greater flexibility in updating the NASS List Frame than was available during the 1997 Census of Agriculture list build process. In 1997, the complete incoming data from a record with the 'best' record source was added to the frame. NASS's resolution review system gives reviewers the capability to build records by using a portion of the information from one source and another portion of the information from another source. For example, if the name information was more complete on one outside list but the address information was more complete on another, the combined best information would be stored on the NASS frame. [2]

Measuring Initial Duplication on the List Frame
Merging the NASS and Census Frames into one combined frame provided many advantages over maintaining two separate frames. However, there was concern that merging the two frames also increased duplication levels. Research was conducted to assess the amount of duplication following the merge. The research was done in Idaho, Missouri, and Texas using NASS's January 2000 Cattle Survey. An attempt was made to identify all detectible duplication among cattle records sampled for the three States. Potential duplicates were reviewed by personnel in the Idaho, Missouri, and Texas Field Offices. Net change in survey indications adjusting for the additional duplicates was derived. Duplication levels for the three States were extrapolated to obtain overall duplication levels for each State's List Frame. The duplicate records from the Cattle Survey expanded to entire list duplication levels that were comparable to levels found in previous NASS studies of list duplication. Duplication levels among active farm records were generally less than .5%. The study suggested that

although there still was duplication on NASS's frame, merging the two lists did not result in a large increase in duplication levels. [3]

Screening Potential Farms

The process of building NASS's List Frame results in a large number of records with unknown farm status being added to the frame. According to available time and resources, States Office personnel attempt to contact as many potential farm records as possible to determine whether or not they actually are farms. Typically, a questionnaire is mailed each year to all potential farms. Follow-up phone calls and/or mailings are often made to questionable operations and nonrespondents. Agricultural data are collected for operations that are involved in agriculture. The efforts to obtain information from potential farms are usually directed by each State Office. However, a large scale national screening operation (Farm Identification Survey or FIS) was conducted at the US level just before the 2002 Census. In March 2002, all potential farms (591,288 records) were pulled from the NASS List Frame. Each record was mailed a short questionnaire with a series of yes/no questions and one question requesting a categorical indication of sales. Nonrespondents were sent a follow-up questionnaire approximately 6 weeks after the initial mailing. A second extract of new potential farms added to the List Frame after the initial pull was pulled in May 2002. This extract included 568,692 records. These records were mailed the same questionnaire. However, nonrespondents were not mailed a follow-up questionnaire. The efforts in determining the true status of potential farms contributed to a more efficient Census Mail List and aided in the unduplication process. Quite a few duplicate records were identified during the data collection phase of the screening process. As duplicates were identified, the best record was retained on the NASS List Frame and the other record(s) were dropped.

Removing Duplication Before the Census Mailout

During the spring and summer of 2002 effort was spent preparing records that would be included in the Census Mail List. State Office personnel worked to improve name and address quality. They also worked to remove duplication both within their State and across the U.S. Each State's List Frame was unduplicated using probabilistic record linkage techniques. The processing and review were done just before the Mail List was pulled in the summer of 2002. Unfortunately, all States were not able to complete their review before the Mail List was pulled. Some States completed the review after the Mail List was pulled. Duplicates identified after the Mail List was pulled were mailed Census questionnaires. However, the records were marked so that their data were not duplicated in the final estimates. Beyond removing duplication within each State, an attempt was also made to identify duplication across States. A procedure was run that identified records with the same SSN, EIN or phone number across States. These potential duplicates were also reviewed by field office personnel. Again, not all States completed this review before the Census Mail List was pulled.

The Census Mail List was pulled on August 31, 2002. The final list included just over 2.8 million records. These records can be broken down into 1.8 million records thought to meet the NASS farm definition and one million potential farm records.

Identifying Duplicates after Data Were Received

As Census questionnaires were returned and analyzed attempts were made to identify and remove duplication. Respondents receiving multiple questionnaires for the same operation were instructed to return all the questionnaires in the same envelope. Clerks, checking in returned questionnaires, were careful to look for and identify potential duplicates. Some respondents phoned the Census Call Centers to report duplication. These duplicates also were identified and removed. New questions were added to the 2002 Census of Agriculture asking for additional names and/or addresses associated with an operation. The information reported for these questions was very helpful in determining whether potential duplicates were separate operations. Approximately 50,000 duplicates were identified and removed through Census processing. It was expected that the duplication rate among records on the Census Mail List would be higher than the duplication rate among active farm records on NASS's List Frame. A large number of the Census Mail List records were records whose farm status was unknown. Many records were identified as potential duplicates to existing farm records prior to the Census Mailout. However, it was unknown whether the records were farming individually or in partnership with existing records. These records were included in the Mail List in an attempt to have complete coverage of the farm population.

Like the 1997 Census of Agriculture, a three-phase duplication edit process was included as part of the 2002 Census of Agriculture. The purpose of the 2002 duplication edit was to identify duplication among the 2002 in-scope Census records after the data were received. The first phase for 2002 was essentially the same as the first phase of the 1997 duplication edit process. It involved identifying and reviewing pairs of in-scope records assigned a link, at some point during the data review process, indicating an in-scope to out-of-scope relationship. Again, a listing of potential duplicate records from this phase of the edit was generated and reviewed by NASS field office staff. The listing contained 1,196 potential duplicates.

The second phase of the duplication edit involved identifying potential duplication based on similar name and address information and reported data. The Within-State Census Unduplication was processed on a State-by-State basis as States reached a 75% response rate. Potential duplicates from the second phase of the edit process were identified through two independent record linkage matches. Links identified in these two matches were combined to make up the final set of potential duplicates that State office personnel reviewed.

The first match linked records based on name and address information. All records with the same SSN, EIN, phone number, or address were brought together as possible duplicates. Additionally, records with very similar name and address information were brought together as potential duplicates. The parameters for this match were set such that the only links output were those links with the highest probabilities of being actual duplicates.

The second match linked records based on reported data items. Any records within the same reported principal county that had five or more data items with a value greater than zero that were exact matches were output as potential matches. Furthermore, records within the same principal county with more than five data items within close proximity of each other were also output. The percent difference in data items was used to calculate a weight representing the likelihood that the items are the same. Thus a record that reported 999 acres would get a positive agreement weight when compared with a record reporting 1000 acres. However a record reporting 4 acres and another record reporting 5 acres would not get a positive agreement weight. As noted previously, links from the two Within-State record linkage projects are combined to make up the final data populated to the resolution database. The Within-State duplication processing identified a total of 15,642 potential duplicates. Upon completion of the review process by State office personnel, approximately 3,500 true duplicates were identified and removed.

The third phase of the duplication edit identified potential duplicates for in-scope Census records across all States. The purpose of this phase was to identify potential duplicate records across State borders. All in-scope Census records were processed in one batch at the same time. The timing of this processing occurred after the majority of the States reached a 75% response rate. Unlike the Within-State Census Unduplication, the processing did not include a probabilistic Record Linkage approach. Rather, listings were produced by independent matching of SSN, EIN, and phone number with a SAS program. Each State had an average of 15 potential Cross-State duplicates for review. Only a small portion of the potential duplicates were true duplicates.

Summary

As part of the 2002 Census of Agriculture, a great deal of effort was placed on removing as much duplication as possible. Attempts were made to eliminate duplication during the Mail List build processes. Duplication was also eliminated as part of the screening and unduplication processing before the Census Mailout. Finally, duplicates were identified as data were received and analyzed. Despite multiple efforts to identify and remove duplication, undetected duplication will still exist upon the completion of the 2002 Census of Agriculture. To learn more about the undetected duplication, a Classification Error Study (CES) will be conducted at the conclusion of the 2002 Census of Agriculture. One purpose of this study will be to obtain a measure of the number of farms duplicated or counted more than once by the Census. The study will also help identify improvements NASS can make in the future to reduce duplication. It is hoped that the results of this study combined with other improvements will produce a 2007 Census Mail List with the maximum possible coverage and minimum level of duplication.

Improvements to Remove Duplication from Future Censuses of Agriculture

The procedures used in identifying and removing duplication from the 2002 Census helped improve the quality of the final estimates. Nevertheless, improvements can be made to increase the detection of duplication and to make the process of identifying and removing duplication more efficient. The following list details some areas where NASS plans to focus efforts in an attempt to improve the unduplication process for the 2007 Census of Agriculture.

1. *Identify known independent operations so they do not have to be reviewed again.* Because of the nature of agricultural operations, having valid independent operations with common identifying operations is possible. For example, two separate operations may use the same mailing address. NASS is considering implementing a system where records would be marked as valid separate operations and reviewed once between Census cycles. With the current procedures, certain pairs of records are output as potential duplicates in almost every unduplication procedure. Time and energy are wasted in reviewing the same records multiple times.

2. *Improve the matching parameters*. NASS is continually working on improving strategies to define the parameters used for its record linkage projects. As more record linkage experience is gained, new ideas will surface which will help make the linkage procedures more accurate and efficient. NASS constantly struggles to achieve a balance between reviewing the fewest records possible and maintaining low linkage error rates.

3. *More consistent review instructions.* The decisions made by NASS's field office personnel during the clerical review process are not always consistent. For example, a person in one office may call a link between two records with the same address but different names a match where someone else in another office may call the same type of link a nonmatch. NASS would like to develop a set of clerical review guidelines to encourage more consistency between State offices. It is hoped that developing these guidelines will reduce the review time as some review rules can be automated as part of the linkage programs.

4. *Enable State office personnel the resources necessary to complete the unduplication review before the mailout.* NASS has a valuable resource with the knowledge of personnel in each of its field offices. Their insight in reviewing potential duplicates results in a higher quality end product. Before the 2002 Census mailout, potential duplicates were identified and made available to field office personnel. Unfortunately, many States did not have the time or resources to review the potential duplicates before the Mail List being pulled.

5. *Enhance the procedures to link records based on reported data.* Part of the process used to identify potential duplication among in-scope 2002 Census records involved linking records based on reported data values. The logic used to do this match in 2002 was identical to the logic used in a similar match done as part of the 1997 Census of Agriculture. For future Censuses, it may be advantageous to include a match on data items as part of the unduplication effort prior to the Census Mailout. Research also can be done to improve the logic used to link records on data items. Some variables used to link records in 1997 and 2002 were not independent. For example, three variables used to link records dealt with farmland. They were land owned, land in the farm, and land in the principal county. It was very common for an operation to report the same value for all three variables. Records that did report the same value for all three only needed to have the same data for two other items to be output for review. This led to many potential duplicates being output that probably did not need to be reviewed. The logic used for the 1997 and 2002 matches only linked records based on reported data items. More duplication may be detected if edited and/or imputed values are also included in the match.

6. *Use probabilistic record linkage on Cross-State duplication check.* Probabilistic record linkage techniques are not currently used to identify Cross-State duplication. NASS is considering implementation of probabilistic record linkage to identify potential duplication across States.

References

[1] Fellegi, Ivan P. and Sunter, Alan B. (1969). *A Theory for Record Linkage.* Journal of the American Statistical Association. 64:1183-1210.

[2] Broadbent, Kara and Bill Iwig. *(1999) Record Linkage at NASS Using AUTOMATCH.* 1999 Federal Committee on Statistical Methodology Research Conference: Complete Proceedings. 2: 595-604.

[3] Anderson, Carter and Kara Daniel. (2000) *Evaluation of the Impact of List Duplication on the January 2000 Cattle Survey Indications*. United States Department of Agriculture, Census and Survey Division, Sampling Branch.