RECORD LINKAGE and AUTOMATIC MAINTENANCE ACTIVITIES

Holly Smith, Kara Daniel, Denise Abreu, Stan Hoge, Bill Iwig
United States Department of Agriculture, National Agricultural Statistics Service
1400 Independence Ave, S.W., Washington, D.C. 20250

Abstract

The National Agricultural Statistics Service (NASS) mission is to provide accurate, reliable, and current agricultural statistics in a timely fashion. In order to achieve this goal, NASS needs a reliable and efficient list sampling frame of agriculture producers in the United States. In 1997, responsibility for the Census of Agriculture was transferred from the Census Bureau to NASS. In addition to the Census, NASS' annual survey program has grown to include economic and environmental statistics as well as production agriculture statistics. These growing responsibilities created the need to explore a way to more efficiently and effectively handle list maintenance activities.

The paper provides an overview of NASS and its procedures for utilizing record linkage techniques for list building and list maintenance. It will briefly describe the way NASS reviews records and some of the features of its record linkage application. It will discuss current record linkage projects being utilized for maintaining variables such as gender, race, and ethnicity on the list frame and the use of record linkage to identify and resolve duplication.

Key Words: maintenance updates, record linkage, List Frame

Introduction

The National Agricultural Statistics Service (NASS) is responsible for the publication of agricultural data. Its mission is to provide timely, accurate, and useful statistics in service to U.S. agriculture. NASS has an extensive annual survey program to collect agricultural related data with Field Offices (FO) in 44 States. In recent years, the survey program has expanded to include the publication of more economic and environmental related data. In addition, the responsibility for the Census of Agriculture was transferred to NASS in 1997. The Census of Agriculture is conducted every 5 years, the next one being in 2007.

One of the most important aspects of an effective survey program is a database of names, addresses and data items from which survey samples are drawn. NASS refers to this database as the List Frame. The NASS List Frame consists of names and addresses along with other control data variables such as corn acres, soybean acres, number of milk cows, number of hogs, etc. Demographic information also resides on the List Frame. Examples of this type of information are: gender, race, ethnicity, age, etc. All this information needs constant review and updating since farm operations change on a frequent basis.

The List Frame is never complete. NASS compensates for the incompleteness of the List Frame by comparing the List Frame to a complete Area Frame. The Area Frame accounts for all land within a State by dividing land acres into segments based on how the land is cultivated. Names and addresses associated with sampled Area Frame segments reside in a database similar to the List Frame. Since the complete Area Frame is used to measure the incompleteness of the List Frame it is critical that both frames are kept separate and information on one frame may not be used to update information on the other frame.

Both frames require constant maintenance and consume many resources.

NASS tries to keep its List Frame as up-to-date as possible by constantly obtaining new agricultural lists from various sources. Information from these list sources is used to update data on existing List Frame records and to add new potential agricultural operations to the List Frame. Large outside list sources are typically matched to the NASS List Fame using the NASS record linkage system. These outside list sources are processed and reviewed on a State by State basis. Before the Census of Agriculture was transferred to NASS, data on most outside list sources consisted of names, addresses, phone numbers, social security numbers (SSNs), employee identification numbers (EINs), and information on crops or

livestock. There were not many outside list sources that contained demographic variables such as race, gender, and ethnicity. It is more common for outside list sources currently being processed through record linkage to have these variables.

List Frame maintenance activities are highly expensive both in human and equipment resources. The NASS record linkage system was designed such that lists could be accurately matched to one another while minimizing human resources. Furthermore, the system was designed such that additions and updates to the List and Area Frames could be made as efficiently and effectively as possible.

This paper will discuss different techniques employed to automate List Frame maintenance activities. It will describe the process through which records are reviewed and discuss the way in which record linkage techniques are used to identify duplication.

Background of NASS and Agriculture

The number of farms in the United States has been dwindling for the past 10 years. Currently, there are an estimated 2 million farms in the United States. The advance of technological breakthroughs has allowed farmers to become more automated. The increase in competition from other countries has pushed many farmers out of agriculture. The trend lately is for large farms to get larger which is cutting out the medium farms. Additionally, there are a growing number of small farms mainly consisting of part-time or hobby farmers. NASS defines a farm as a person or establishment with $1,000 value of sales of agricultural products annually or potential value of sales of $1,000. Many individuals who have a few cows or a few horses do not consider themselves farms even though by NASS definition they would be classified as farms. New outside lists are gathered and matched against the List Frame in an effort to build a list for the Census of Agriculture and the NASS Survey program.

The agriculture sector is complex and shifts on a daily basis. The NASS vision for its List Frame is an operator dominant view. An operator dominant view has the target record as the operator, not the operation. There are a few cases where the operation is the target record. Those situations are specific to operations that do not change and are stable. They are typically large operations that have been in business for many years. Using the operator as the target creates complex relationships that are difficult to capture and maintain on the List Frame. Specifically, there are situations where one operator may be involved in many operations or many operators are involved in one operation. Reflecting this situation on the List Frame can be a difficult process. Additionally, operation names, partnerships, and target operators change on a daily basis.

NASS and Record Linkage

NASS began utilizing its current record linkage system over 9 years ago. The NASS system was built using AutoStan and AutoMatch (formerly sold by MatchWare Technologies) as its base for the standardization and matching of records. These software programs were developed using the probabilistic record linkage techniques proposed by Ivan Fellegi and Alan Sunter in their 1969 JASA paper. NASS developed front and back end companion products to assist in setting up record linkage match parameters and reviewing results.

NASS begins a typical record linkage project by obtaining an outside list source. This list is then transformed into a standard fixed field ASCII text file. Data in this file are formatted such that they meet the standards of the List Frame. Individual names are transformed to signature format. Variable length restrictions are imposed so that the length of fields going into record linkage match the length of the corresponding fields on the List Frame. For example, name and address fields are limited to 30 characters on the List Frame. So the name and address fields are reformatted to 30 characters. A list identification number is generated for each record in the outside source list. This allows the processor to easily identify the record before and after the match is run. Each outside source list is assigned a record source code. The record source code is an indicator of the outside list source. Each outside source record is also assigned a status code. Examples of the status code are: known farm, potential farm, or non-farm record. The status code is continually updated based on information received about the record.

Before processing a record linkage project, a fixed field ASCII text file is pulled from the NASS List Frame database. The layout of this

file is identical to the layout of the outside source ASCII file. The List Frame extract typically contains all records on the frame, including known farms, potential farms, and non-farms.

After the outside source and List Frame fixed field ASCII text files are created, a SAS program is run to determine if the outside source cities and ZIP Codes correspond to the United States Post Office standard cities and ZIP Codes. If a postal standard city and ZIP Code cannot be found, a report is generated noting the city or ZIP Code information is incorrect and should be updated before the match is processed. The SAS program also verifies that telephone numbers, SSNs, and EINs meet certain validity standards. If they are not valid, a report is generated. Once the verify program is finished and the file is free of all errors, a standardization process is run.

This standardization process is run for both the outside source and the List Frame files. The standardization process parses out the names and addresses into their component parts. For example, a person name could be parsed out into a prefix, first name, middle name, last name, last name suffix and title. During the standardization process, input name and address components are replaced with standard values. This standardization removes the effect of common nicknames and spelling variations. It also ensures that like information is compared during the match process.

NASS utilizes AutoMatch software to match the outside source and List Frame files. For each project, a set of blocking variables is used to divide the data into mutually exclusive subgroups. Records with common values for the blocking variables are compared during the match process. Records where the blocking variables differ are considered non-matches. For example, one pass may block on the ZIP Code. Records with common ZIP Codes are compared during the match process. Records with different ZIP Codes are considered non-matches. Multiple passes with different blocking and matching variables are run to compensate for inaccuracies in the data. Once the blocks have been determined, values for a series of match variables are compared. If the values agree, a positive weight is generally assigned. If the values disagree, a negative weight is generally assigned. The weights for each of the variables are then summed up to come up with a composite weight. The composite weight is a measure of the likelihood that a record pair is a match.

After a pass is run, a report is generated showing possible links. These links are reviewed in SAS. Linked pairs are sorted according to their composite weights. During the review process, two cutoff values are set for each pass. These are referred to as the upper and lower cutoffs. Any record pairs with a composite weight higher than the upper cutoff are considered matches. Record pairs with composite weights lower than the lower cutoff value are considered non-matches. Record pairs with weights between the two values are considered possible matches. The possible match records need manual review before a final review status will be set. Ideally this review would be done between each pass. However, this review would become logistically infeasible because review of possible matches is done by personnel in the Field Offices. Rather than review possible matches between each pass, NASS combines the results from all passes into one final review. The results are combined using a SAS program. This SAS program brings all related records together into groups which NASS refers to as link groups. A link group contains all match pairs involving the same outside source or List Frame records. Additionally, all records marked on the List Frame as associated with an operation are brought into the group (for example, partner or manager records). Each link group is classified as either a match, possible match, or non-match based on how the pairs that make up the link group were classified during the matching process.

NASS developed its own resolution system for reviewing record linkage projects. After an outside list source has been matched to the List Frame, a record linkage database is populated with the results. This database is independent of the List Frame. However, the resolution system has many features which allow those performing resolution to view and update List Frame information as needed. FO staff does the majority of the resolution work. The employees in the FO have experience with the agriculture in that State and work closely with important State Agriculture contacts. To resolve the link group, FO staff review the records and determine the match status. At times, phone calls are made to verify operating arrangements.

When the FO reviews the link groups, the reviewer goes through the link group to determine which List Frame record best matches

the outside source record. If an outside source record does not best match the first List Frame record in the link group, the reviewer can change the link group number so that the outside source record and the best List Frame record are in the same link group. The reviewer can create up to 10 new link groups (sub link groups) for a particular existing link group. The NASS record linkage system creates a composite record which represents each operation contained in a link group. Once all the records are grouped with their best List Frame record, the composite record is regenerated so that a composite record exists for each sub link group. These composite records are the records used to generate any transaction files for possible name and address, control data, and any other updates that are desired. Additionally, the composite records are the records used to create new add records to be posted to the List Frame. Users have the ability to alter information in the composite record so that it contains information for each operation that is as accurate as possible.

Maintenance Updates

NASS uses record linkage to automatically update the List Frame. This process saves the agency time and resources. Updates prior to the current NASS record linkage system were done manually and were resource intensive. The rest of the paper will focus on describing these update processes.

General List Frame Updates

The List Frame Section processes many record linkage projects each year. FO's are encouraged to gather as many State lists as possible and run them through record linkage to increase coverage of farms and accuracy of data on the List Frame. Headquarters personnel also are encouraged to pursue lists on a national basis and those lists are processed through record linkage as well.

As outside source lists are processed, it is common for the outside source records to have different information than the corresponding information on the List Frame. For example, the outside source record may list a different address than the address stored for the corresponding NASS List Frame record. As FO staff review link groups in the resolution system, they have the option of updating the differing information on the composite record as appropriate. Outside source lists vary on the information they contain. Consequently, the type of updates made to the List Frame also varies depending on the list source. At a minimum, outside source lists contain a name, address, and a farm operating status code. In most projects, these variables are updated for List Frame records. Depending on the outside list source, almost every other variable on the NASS List Frame has the potential to be updated through resolution.

Once resolution is complete, transactions can be generated to post the updates from the composite records to the List Frame automatically. Most of the automatic updates use a feature that creates transaction files for a project. These updates are generated from the resolution database. Each project is divided into subprojects based on certain criteria. Most projects are divided into 9 subprojects according to believed operating status or size. As the review of a subproject is completed, transactions are generated and automatically loaded to the List Frame. Users have the option of selecting all variables or only checking those variables the user wants to update. Users also have the option of creating a listing of the transactions that can be reviewed before posting the updates. This gives FO's the option of reviewing changes before they are posted to the List Frame.

Figure 1 (Update Query Screen) below is the screen used by the FO staff to create the automatic transaction files. The reviewer chooses the variable desired for the automatic updates and selects the generate button and the files are created. The files created are formatted in a specific format unique to the application used to update the List Frame. Transaction files from all 9 subprojects can be batch updated together. There are four variables that are not allowed to be updated automatically. These are grayed out in Figure 1. These variables are important to the operator dominant view of the List Frame. All updates for these variables require manual review and update procedures.

Outside source lists also often contain agricultural operations that are not present on the List Frame. These records are generally non-matches in the resolution system. However, they also can be identified when records linked to each other in the resolution system are unlinked.

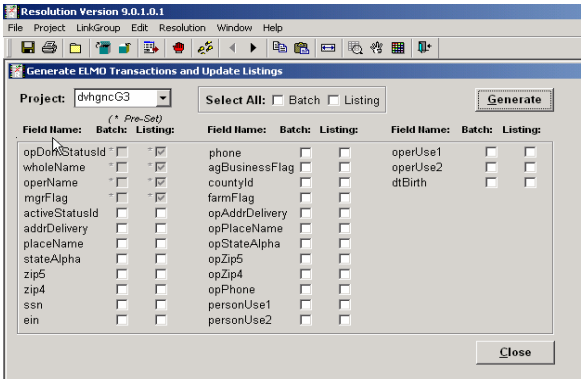

Figure 1: Update Query Screen

After resolution is complete, all outside source records that are not linked to a List Frame record are added as new records. It is usually unknown whether these records meet the NASS farm definition. Hence, most new records added to the List Frame through record linkage are added with a status code indicating that data needs to be obtained to determine the record's true operating status.

Telephone Updates

NASS processes two specific projects to target telephone changes. These projects are usually run multiple times each year with timing targeted so that telephone updates will be posted prior to any large telephone calling effort. It is much more cost effective to bulk update telephone numbers through record linkage than to have clerks try to look up records with missing telephone numbers on the internet. Both record linkage projects involve matching all List Frame records with missing or invalid telephone numbers to an outside source list. For the first project, the List Frame records are linked to a national telephone database obtained from InfoUSA. For the second project, the List Frame records are linked to records on the United States Department of Agriculture Farm Service Agency's (FSA) List. Telephone numbers in the InfoUSA database are generally more up-to-date than those on the FSA List. Hence, the InfoUSA match is typically run prior to the FSA match. For both projects, cutoffs are set such that only matches and non-matches are identified. For match groups, the outside source record's telephone number is automatically posted to the List Frame. There typically is no possible match resolution in either project. No action is taken with non-match records.

SSN/EIN Updates

Most records on the FSA list have either a SSN or EIN. NASS runs two companion record linkage projects designed to update its List Frame with FSA SSNs and EINs. Past work with FSA records has shown that the SSNs and EINs on a record do not always correspond to the name on the record. It is not unusual for a name to have a family member's SSN or EIN. Hence, FO's have the option of reviewing these changes before they are posted to the List Frame. For both projects, the only link groups identified are those with weights high enough to be classified as matches. No non-match records are output or reviewed.

Post Office Return Address Updates

NASS is running a new record linkage project this year. This project will match List Frame records with status codes indicating they were Post Office returns and missing address records to records on the FSA List. Links made through this match will be reviewed and good FSA addresses will be used to update the List Frame.

Control Data Updates

Many times, outside source lists contain control data elements that can be used to update information on the List Frame. Control data items include total acres, total cropland, corn acres, soybean acres, number of cattle, number of hogs, etc. For example, one match done each year is the Agricultural Marketing Service (AMS) dairy project. A list from AMS is acquired with names, addresses, AMS ID numbers, and milk production. Given this information, an estimated number of dairy cows can be derived. The records for each FO are run through record linkage separately. FO staff can review and make the needed name and address updates. Additionally, when the project is complete, the derived number of dairy cows is posted to the List Frame and is used to sample cattle each year. This match helps to identify new dairies the FO does not already have on the List Frame.

Demographic Updates

In the past, NASS made very few demographic updates. Demographic updates are becoming more common, especially with the added responsibility to conduct the Census of

Agriculture. One upcoming demographic project will link FSA records with minority race codes on the List Frame. FO staff will have the option of reviewing cases where the FSA and NASS have differing race codes for a record. The FSA race codes will be updated on NASS records that do not have a race code. Furthermore, FSA minority records thought to be likely farms that are not on the List Frame will be added as new potential farm records.

Last year, NASS ran a project designed at identifying aquaculture operations that needed to be included in the Census of Aquaculture. Several aquaculture lists were obtained which contained NAICS (North American Industry Classification System) codes. Following resolution of these lists, the aquaculture NAICS codes were posted to the NASS List Frame. These NAICS codes were used as one of the criterion for identifying records to be included in the Census of Agriculture.

Active Status Updates

Every record on the List Frame has a status code associated with it. This code indicates the believed agricultural operating status for a record. The active status of a record determines its ability to be included in a population during the sampling cycle. Only records with certain status codes are considered for sampling purposes. During resolution, non-active records can be linked to outside source lists. When this situation occurs, the active status of the List Frame record is changed to reflect the possibility the record may be in business. These records are then selected for an agricultural screening survey to determine business status. For example, if a record with an out-of-business status code is linked to a current record on an organic grower outside source list, the status code on the List Frame record would likely be changed from an out-of-business code to a code indicating the record is a potential farm.

NASS runs one record linkage project each year with the specific purpose of updating status codes. This project is the Social Security Death Match. Each year, NASS obtains files from the Social Security Administration's (SSA) Death Master File. This file contains records of SSA payments for deceased individuals. The goal of this match is to identify active farm records whose operator is now deceased and also to update the status of out-of-business farm records to a deceased status code. Active farm records do not have their status code automatically updated to an inactive status. Rather, these records are marked for follow-up work and surveys are conducted on these records to determine whether the farm is still active. If the operator is in fact deceased, an attempt is made to learn if someone else took over the operation. Only out-of-business List Frame records matched with death match records have their active status automatically updated.

Record Source Id

Each record in an outside source list has a record source identification number attached to it. This source indicates the original source from which each List Frame record was added. The record source is helpful because it often gives the indication as to a record's expected type of operation. Record sources are also useful because they can be used to perform analysis evaluating the effectiveness of a list. When the resolution process is complete, records in the non-match category are added to the List Frame. During and after completion of the 2007 Census of Agriculture, all list sources will be evaluated to determine which ones were the better sources of identifying active farms. These lists should be acquired in the future before the next Census of Agriculture. Evaluation of list sources will be important in building the Census Mail List for the 2012 Census of Agriculture and the NASS annual survey program.

Record Linkage Match Year

The record linkage match year is a field on the List Frame that identifies the last time a record was matched to an outside source list in a record linkage project. This field is one of many sources that help NASS assess the likelihood that records are still involved in agriculture.

Linkage Information

NASS stores a variety of linkage information on its List Frame. Information is stored linking records with common characteristics. For example, two separate operations that have the same address often have a link stored on the NASS List Frame. This link ensures that both records will always appear in the same link group in future record linkage projects.

Links are also stored which connect NASS List Frame ID numbers to ID numbers on outside source lists. Storing these outside source numbers is helpful because the ID numbers can be matched on future lists that are re-run on a periodic basis. Having the outside source ID available as a matching variable greatly reduces the record linkage error rates as well as the cost associated with reviewing a project.

A good example of a project that utilized linkage information is the project discussed above in the control data section. In that section the AMS dairy project was described. AMS ID numbers are stored on the NASS List Frame and used each year to bring records together.

Record Level Comments

Both the NASS List Frame and the record linkage resolution database allow users to store comments for records. While FO personnel are reviewing records, comments pertaining to a link group can be reviewed and updated. Comments specific to resolution can be posted to the resolution database, but not the NASS List Frame. FO staff find these comments helpful when processing other record linkage projects, performing manual updates, or identifying duplication on the List Frame.

Identifying List Duplication

Record linkage is used to detect duplication among NASS List Frame records. Each FO usually goes through a List-to-List unduplication process each year before the start of the sampling cycle. This unduplication effort involves identifying potential duplicate agricultural records on each State's List Frame. All active farm records and potential farm records from the List Frame are extracted and used in the matching. Several passes are used to help bring together records that could be matches. No matches are identified in this project. Rather all link groups are classified as possible matches and reviewed. Field Offices review the possible match link groups and make updates to the List Frame as needed. The potential duplicate review is one of the more difficult record linkage projects and often requires a contact be made with the farm operator. Automatic updates are generally not run for this project.

In addition to the List-to-List unduplication effort, an Area Frame to List Frame duplication project is run each year. The base area survey is conducted each June. Prior to the start of the June Area Survey, an Area-to-List duplication effort is done matching records on each State's Area Frame sample to their List Frame. This is done to determine which area sample records are also on the List Frame. Each June, an Area Survey is conducted to determine the names, addresses and control data information of all the operators in each segment. The information is added to the Area Frame after data collection. After the information has been added, a post Area Frame to List Frame duplication effort is done to link any new Area Frame records to existing List Frame records. Area Frame records are linked to List Frame records because Area Frame records that are not on the List Frame are used in future surveys.

Conclusion

NASS continues to refine its record linkage resolution system to meet the changing needs of the agency. NASS utilizes its record linkage system to automatically perform many maintenance updates. There are several new projects related to creating new features for the resolution system. NASS is currently working on the following:

- Create the option to generate the update transaction files for more than one subproject at a time. The record linkage system can only generate transactions for one subproject at a time. At times, it would be beneficial to give the Field Office personnel the option to process two or more subprojects at one time.
- Add more variables to the record linkage database. Presently, there are a limited number of variables record linkage can handle when processing a project. There is a need to add new variables like gender, race, ethnicity, more control data variables, email, additional phone numbers, and several more. NASS is working on adding these variables to the record linkage database to make resolving records easier and increase the number of automatic updates available.
- Mark records with an additional flag indicating the need for verifying information with an operator. The flag and operator's name and address information would be read out of record linkage using the transaction process to

generate an ASCII file to be used at the Field Office discretion.

- Develop a way to allow the user to create a resolution screen layout with the variables desired to review a record linkage project. At the moment, there is not a good way to switch variables around within the resolution screen except for dragging the variable to the desired location. It would be useful to have an application that can be custom driven based on the user and the record linkage project.
- Create a button on the record linkage screen to allow users to automatically add control data to records on the List Frame. This would be similar to the current ability of users to add comments to the List Frame from the resolution screen.
- Create an application where users can key outside source records name and address information. This application would then create a file reformatted to the desired specifications of the record linkage system. Much of the time spent on populating a record linkage project is spent formatting a list to match the format needed to process it through record linkage.

NASS is working on expanding the current record linkage system to handle more variables and automate more of the maintenance processing.

REFERENCES:

Fellegi, I. P. and Sunter, A. B. (1969) "A Theory for Record Linkage", *Journal of the American Statistics Association*, 64, 1183-1210.

Broadbent, K and Iwig, W. "Record Linkage at NASS Using Automatch". (1999), *FCSM Research Conference, http://*www.fcsm.gov/99papers/broadbent.pdf