RECORD LINKAGE AT NASS USING AUTOMATCH

Kara Broadbent, Bill Iwig
National Agricultural Statistics Service, United States Department of Agriculture

Key Words: Record Linkage, List Sampling Frame, AutoMatch

Introduction

The National Agricultural Statistics Service (NASS) gathers information concerning various aspects of agriculture through a system of surveys. The samples for most of these surveys are drawn from a register of farm operators and agri-businesses known as the List Sampling Frame (List Frame). Separate frames are maintained for each state by NASS's State Statistical Offices (SSOs).

Maintaining a current, accurate List Frame is a time-consuming and expensive process. As new list sources are obtained, duplication is removed, and the list is matched against the master List Frame file using probabilistic record linkage techniques. Over the past couple years, NASS has developed a new record linkage system with AutoMatch as the core. AutoMatch was developed by MatchWare Technologies using the record linkage theory proposed by Ivan Fellegi and Alan Sunter in a 1969 JASA paper [1]. NASS's new system compliments the AutoMatch system by providing front-ends for defining the match parameters and back-ends to facilitate the process of reviewing records and updating and adding new information to the List Frame.

This paper will give a general overview of the record linkage techniques used by NASS and will describe the system developed for review.

General Methodology

The List Sampling Frame is a register of all known agri-businesses and farm operators. To keep this list as current and accurate as possible, it must be continually updated with information from several different agricultural list sources. These new list sources do not always have common layouts or formats. One list may have names in signature format (John Smith) while the other may have names in surname on the left format (Smith, John). Furthermore, the incoming lists do not always contain the same data fields. They generally will contain name and address information, but vary on the presence or absence of other fields such as social security number and phone number. The new record linkage system is flexible enough to adapt to the special situations presented by each list. A series of front end screens allow users to specify the format and fields present in the incoming new source lists. Default format parameters have been set up for incoming lists that are obtained on a periodic basis.

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

Once the format of a new list source has been defined, the incoming file will be transformed into a standard layout for matching. New source lists are matched against an extracted file of records from the List Frame on a state by state basis using AutoMatch. AutoMatch links records using the probabilistic record linkage techniques proposed by Fellegi and Sunter in 1969. It assigns a weight to each component of the records being compared. These component weights are then summed to calculate an aggregate weight for the record pair. The aggregate weight represents the probability that the record pair is a true match. The aggregate weight is compared against two thresholds, or cutoffs, to classify each case as a match (above the upper cutoff), nonmatch (below the lower cutoff) or possible match (between the upper and lower cutoff).

AutoMatch is run using a series of passes. A linkage identified in an early pass may contain some of the same records as a linkage identified in a later pass. For example, a new source record may match a List Frame record in pass 2 and a different new source record may match the same List Frame record in pass 7. A merge program was written in SAS to combine the records from both AutoMatch linkage groups into one conglomerate link group. The merge program also brings together any additional List Frame records that are related to the same operation (such as partners) based on common values in various ID fields. Multiple List Frame records that are involved with the same operation (for example, a primary operator and partners) are linked together by a common value in the operator ID field. Similarly, multiple operations that involve the same person are linked together by a common value in the person ID field. Additionally, List Frame records may have other common characteristics (such as two different operations at the same address) and be linked to one another based on a general List Frame linkage Id field.

Each conglomerate link group formed by the merge program is classified as a match, possible match, or residual group. If any individual linkage between records within a link group was classified as a possible match, the entire link group is classified as a possible match. Link groups that contain List Frame records from multiple operations are also classified as possible matches. These groups are classified as possible matches because they generally represent complex operations that need to be reviewed before they are used to make updates to the master List Frame. After running the merge program, the link groups are populated into a resolution database where possible match resolution is performed by state office users who are familiar with the farm operations in their state. The NASS record linkage system was designed to make the record linkage process as simple and user friendly as possible.

Matching

The AutoMatch software links records both within and between lists. AutoMatch brings records together through a series of passes where different blocking variables, matching variables, and thresholds are used. Throughout the match process the master List Frame file is treated as a reference file. All List Frame records are included in each pass. However, once a new source record is linked to a master List Frame record, it is excluded from all subsequent passes.

Ideally, the possible matches from a given pass will be reviewed before the next pass is run. This is done because residual records from one pass are used as the new source input file for the subsequent pass. Reviewing the possible matches between each pass allows a new source record that was brought together with a reference file record in an early pass as a possible match and unlinked during possible match resolution to have the possibility of being linked to another reference file record in subsequent passes. The following scenario illustrates the consequences of performing the clerical review between each pass as opposed to a single clerical review after running all passes (See **Figure 1.**).

Suppose that Record 10 from File A is a true match to Record 5 from File B. However, in pass 1, Record 10 from File A is linked as a possible match to Record 12 from File B. If a clerical review is performed before pass 2 is run, Record 10 from File A would be unlinked from Record 12 from File B. Record 10 would then be included in the new source records going into pass 2 and correctly linked to Record 5 from File B. However, if a clerical review was not performed between passes 1 and 2, Record 10 from File A would be considered a 'linked record' and would not be included in pass 2. During the final clerical review it would be unlinked, and its final match status would be a nonmatch. It would be a false nonmatch error because in reality Record 10 should have been a match to Record 5 from File B.

Figure 1. – Separate Clerical Reviews between each pass or one Clerical Review after all passes.

Record 10 from File A Record 5 from File B	True match
Pass 1:	
Record 10 from File A Record 12 from File B	Brought together as possible matches
If a clerical review is done between passes 1 and 2:	
Pass 2:	
Record 10 from File A Record 5 from File B	Brought together as definite matches
If a clerical review is not done between passes 1 and 2:	
Pass 2:	
Record 10 from File A will not be included in pass 2 and will be a nonmatch in the end.	

Because of the number of people and the number of steps involved in reviewing possible matches between each pass, management of the resolution process can often become quite complex. For this reason, research was conducted to determine the consequences of performing one clerical review after all passes have been run rather than a clerical review between each pass.

A pair of files where the true match status of each record was known was used to conduct the research. The files consisted of a new source file with 24,843 names and addresses. This file was matched against a reference file with 117,246 records. With the particular match parameters used for the research, it was found that only 116 additional records were correctly classified as matches when the clerical review was conducted between each pass as opposed to a final clerical review after all passes. This is less than 1 percent of the total number of records. It was decided that the number of matches missed by performing a single clerical review after all passes was small enough that the reduction in staff resources needed to performing a single clerical review after all passes had been run outweighed the disadvantage of missing a few matches that would have been identified if a clerical review was performed between each pass. Thus, in the NASS's new record linkage system, records are matched using AutoMatch, the output files for all passes are then combined, and only one clerical review is performed after all passes are run.

The results of the test comparing one clerical review after all passes to separate clerical reviews between each pass are highly dependent upon the blocking variables, matching variables, and threshold values that were chosen. The matching scenario was developed so that pairs of records that have a high probability of representing the same operation were identified early in the record linkage process. Records were linked on more reliable matching variables in the early passes. Most of the links in the early passes were definite matches. The later passes were matched using less reliable matching variables. Most of the links in the later passes were possible match links. Identifying the records that had a high probability of representing the same operation early in the matching process helped reduce the number of times that a new source record would be incorrectly linked to reference file record as a possible match and eliminated from the realm of comparison in subsequent matches.

The following table illustrates the match parameters used for a typical match between a new source list and the List Frame. This type of matching strategy is used because it identifies records that are likely matches in early passes. The table shows the pass number, the blocking and matching variables, and a sample of the number of matches and possible matches (clericals) for each pass.

Table 1. – Typical Matching Strategy and Counts

Pass	Blocking Variable	Matching Variable(s)	Matches	Possible Matches
1	Zip Code	SSN (Social Security Number)	3306	27
2	NYSIIS of Surname*	SSN	256	1
3	EIN	EIN (Employer Identification Number)	111	14
4	Zip Code	person name and address variables	2205	436
5	Zip Code	operation name and address variables	79	26
6	NYSIIS of Surname	person name and address variables	236	70
7	7 digit phone number	7 digit phone number	120	55
8	Zip Code	person name and address variables	0	535
9	Zip Code	operation name and address variables	0	127
10	NYSIIS of Surname	person name and address variables	0	79

* The NYSIIS code is a phonetic coding scheme used to reduce the effect of different spellings of the same name.

Note that variables involved in passes 4-6 are the same as those in passes 8-10. However, the thresholds used for these passes were different. The thresholds for passes 4-6 were set high enough that the only links identified in these passes are those with high enough weights to be classified as definite matches. The thresholds for passes 8-10 were designed to pick up record pairs that had a lower probability of representing the same operation or the possible matches. The possible matches weren’t identified until the later passes to help reduce the possibility of missing correct linkages of records.

Default matching parameters are being developed for new source lists that are periodically matched against the List Frame. The parameters are developed such that they can initially be used for states. After the initial run, the cutoffs and other values can be altered as appropriate to improve the matching process for each state.

Possible Match Resolution Screens

When new list sources are matched against the master List Frame, there are often a large number of records to review. AutoMatch provides an online clerical review program. However, this program has several limitations and did not sufficiently meet NASS’s needs. A new clerical review system was developed using PowerBuilder. It accesses the data that are stored in a Sybase database.

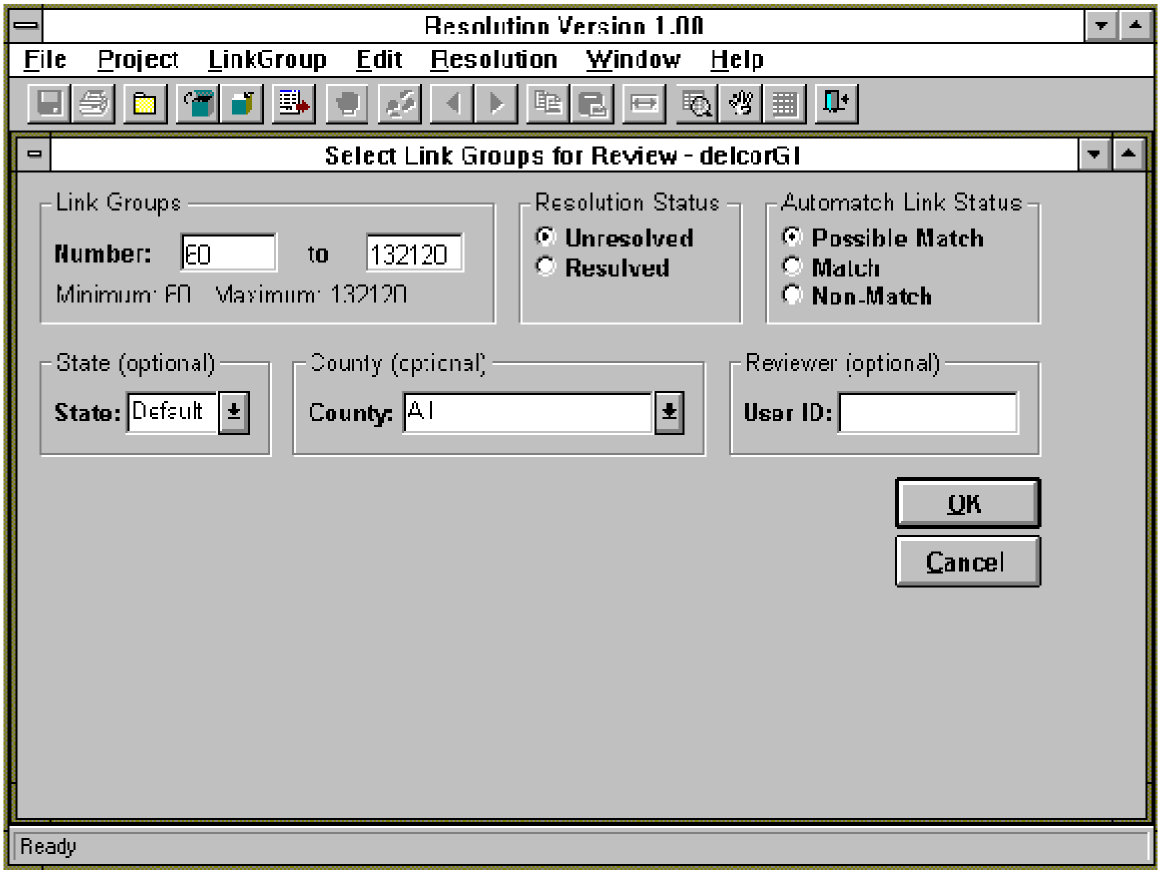
When link groups are populated into the Sybase database, they carry with them a link status based on the match status generated by AutoMatch. This link status will be either a match, possible match, or residual. Each link group also carries a resolution status. The two possible resolution status values are resolved or unresolved. The unresolved link groups are those that still need to be reviewed. Generally all matches and residuals will initially have a resolved resolution status and possible matches will have an unresolved status. At times matches or nonmatches containing records with special characteristics that necessitate review will be initially unresolved. An example would be a link group that contains a List Frame record representing an extremely large farm operation. Marking these cases as unresolved makes it easier to identify groups that need to be reviewed. It also prevents changes from being made to large or complex operations without review.

To make the clerical review task more manageable, the link groups are broken into homogeneous subprojects (see **Table 2.**). The subprojects are designed such that users may concentrate more effort on review of the larger and more complex operations. The subprojects are generally formed by breaking up the link groups by value of sales. The subprojects are hierarchal, meaning that if a record meets the criteria for more than one subproject it will fall into the first subproject it qualifies for. The same person will usually review all the records in a particular subproject. This helps to maintain a consistent review of similar records.

Table 2.-- Description of Subprojects

Subproject	Description
Subproject 1	Operations with addresses in states
Subproject 2	Complex operations (one person involved in multiple operations etc.)
Subproject 3	Operations that are no longer in business
Subproject 4	Value of sales over $500,000
Subproject 5	Value of sales between $100,000 and $499,999
Subproject 6	Value of sales between $50,000 and $99,999
Subproject 7	Value of sales between $10,000 and $49,999
Subproject 8	Value of sales between $1,000 and $9,999
Subproject 9	Value of sales less than $1,000

After logging into the resolution screens, users may select the subproject they wish to work on. After selecting a subproject, the following screen (**Select Link Groups for Review**) will appear. Multiple users may work on the same subproject at the same time.


Resolution Version 1.00
File Project LinkGroup Edit Resolution Window Help
Select Link Groups for Review - delcorGI
OK
Cancel
Ready

Link Groups

Number: 60 to 132120

Minimum: 60 Maximum: 132120

Resolution Status

● Unresolved

○ Resolved

Automatch Link Status

● Possible Match

○ Match

○ Non-Match

State (optional)

State: Default

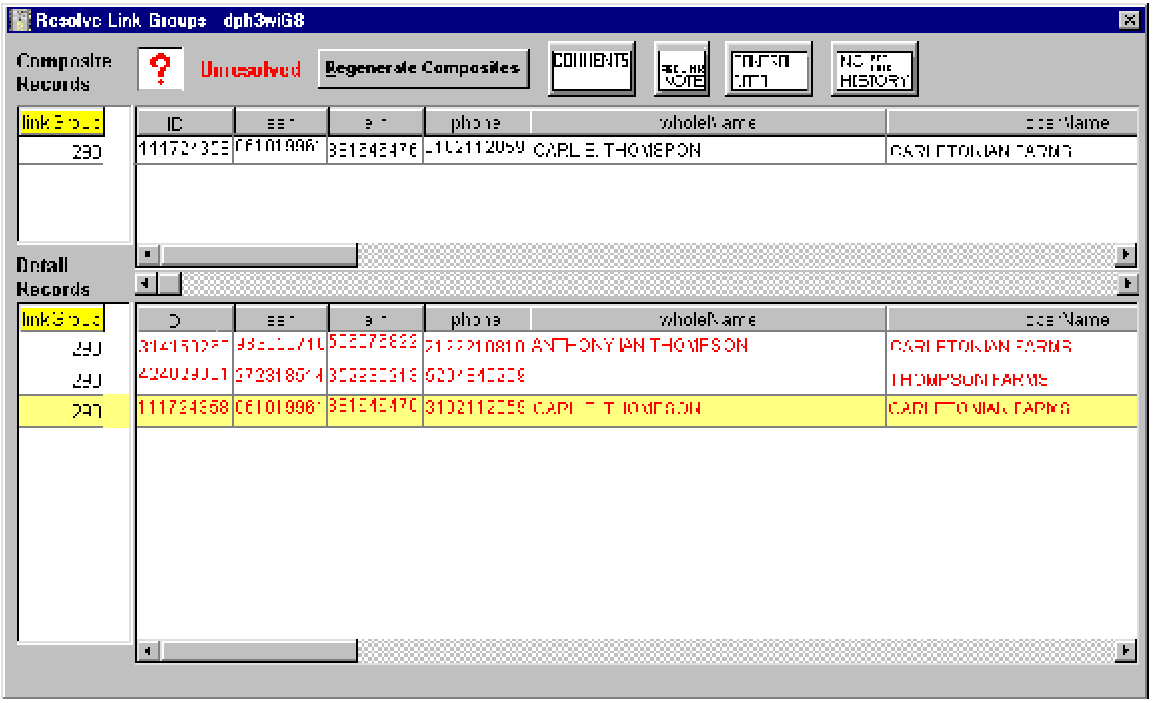
County (optional)

County: AI

Reviewer (optional)

User ID:

The **Select Link Groups for Review** screen controls the link groups which are available for review. Link groups may be selected by resolution status, user ID, AutoMatch link status, county, or state. This screen defaults to show all unresolved possible matches. Once the appropriate parameters are entered, and the OK button is clicked, the program will advance to the **Resolve Link Groups** Screen. The following is an image of that screen.



linkGroup	ID	[illegible]	[illegible]	phone	wholeName	[illegible]Name
230	[illegible]	[illegible]	[illegible]	[illegible]	CARL E. THOMPSON	CARLETONIAN FARMS

linkGroup	ID	[illegible]	[illegible]	phone	wholeName	[illegible]Name
230	[illegible]	[illegible]	[illegible]	[illegible]	ANTHONY IAN THOMPSON	CARLETONIAN FARMS
230	[illegible]	[illegible]	[illegible]	[illegible]		THOMPSON FARMS
230	111724358	06101996	[illegible]	[illegible]	CARL E. THOMPSON	CARLETONIAN FARMS

There are two primary areas in the **Resolve Link Groups** screen. The upper portion of the screen is the composite record area and the lower portion is the detail record area.

The composite record is a single record that represents all records in a link group. It is used to generate transactions to update existing List Frame records and to add new records to the List Frame. The composite record is initially formed by taking the first List Frame record in the link group. If there are no master List Frame records in the link group, the first new source record is used for the Composite Record. If there is only one record in a link group, the Composite Record is the same as the single record. The record used to generate the composite record can be identified because it will have the same ID as the ID in the composite record.

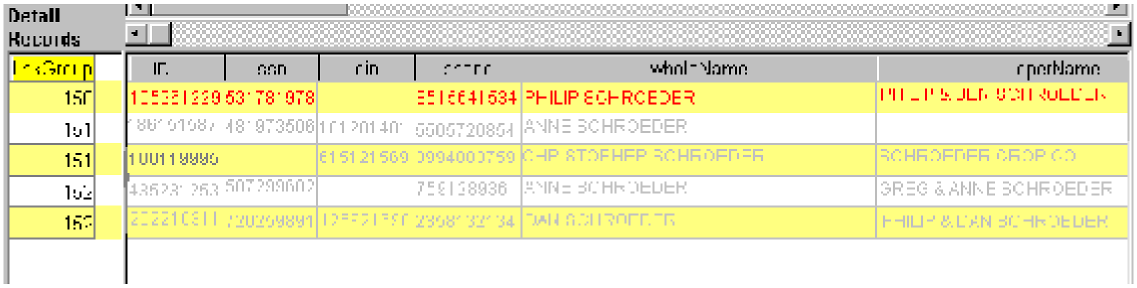
It is important that the composite record is as correct as possible because it is used to update and add records to the List Frame. If a field such as the phone number, SSN, EIN, or Date of Birth, is missing from the record selected for the Composite Record, but present in any of the new source records, the value will be automatically copied from the new source record into the Composite Record. Any data fields that are copied to the Composite Record will appear in green text. This identifies them as fields that will be used to update data on the List Frame. No data will be automatically copied to name or address fields. However, these fields can be edited as needed during the resolution process. Users can edit most fields in the composite records. The only fields that cannot be edited are ID fields. There are several ways a field in the Composite Record can be edited. The four methods are:

1. Copying and pasting data,
2. Dragging data from a field in the detail record, and dropping it a field in the composite record,
3. Deleting fields, and
4. Editing an entire field or part of a field.

As with the fields, like phone number, that are copied to the composite record, fields in the composite record that have been edited will appear in green text. These fields will also be updated on the List Frame.

The detail record portion of the screen displays the original data for all records in the link group along with a link group number for each record. This link group number is a mapping as to which records represent the same operation. The records in the detail record portion of the screen appear with two different background colors. The new source records have a white background and the master file records have a yellow background. Besides different background colors, the records in the detail record portion of the screen have different text colors signifying the match status that AutoMatch originally assigned the link group. Records in blue text were originally classified as matches by AutoMatch, records in red text as possible matches and records in light gray text as nonmatches. No edits are allowed to data in the detail records.

During resolution, a link group may be split up into as many as 10 single link groups by changing the last digit of the link group number field. Each of these new link groups will have a new composite record associated with it. Composite Records for any new link groups that are created will be generated by clicking on the '**Regenerate Composites**' button. This should only be done after the reviewer is satisfied that the link groups are correctly assigned for all records within a group that AutoMatch originally brought together. It is important to note that if any edits were made to the Composite Record, they will be lost when the **Regenerate Composites** button is clicked. The screen below shows an example of the detail record portion of the screen for a link group that was split into three link groups.



Detail Records

lnkGroup	ID	ssn	ein	phone	wholeName	operName
150	125321229	531781978		5516641634	PHILIP SCHROEDER	PHILIP & DAN SCHROEDER
151	180151987	481973506	101120140	5505720854	ANNE SCHROEDER	
151	100119996		615121569	3994003759	CHRISTOPHER SCHROEDER	SCHROEDER CROP CO
152	435231253	507299602		759138936	ANNE SCHROEDER	GREG & ANNE SCHROEDER
152	222210311	120259891	125521551	2368132134	DAN SCHROEDER	PHILIP & DAN SCHROEDER

The ‘resolved/unresolved’ button in the top center of the screen is used to indicate link groups that need to be reviewed and those which are complete. Once satisfied with the decisions made for a particular link group, the resolve button in the top center of the screen can be clicked to mark the record as resolved. As stated earlier, updates and adds are generated for the master List Frame based on the composite record. These adds and updates only occur for link groups with a resolved status. Thus, if for some reason, a user is unable to determine the relationship between records in a link group, he or she may leave the link group unresolved and no transactions will be generated for the master List Frame.

The resolution screens were designed with a number of different features to allow users to customize the review screens to meet their individual needs. Some of the features are:

- The variables can be resized and reordered.
- The variable name headers can be changed in some of the columns.
- If a change in variable order or size is made to the detail records, a button can be pressed to synchronize the order and size of the variables in the composite records.
- “Hot keys” are defined so that users may use either the mouse or the keyboard to move within and between link groups.
- Users can enter comments or view existing comments on the master List Frame or within the resolution program.
- The system records the user ID and time that a change was made to a link group. This is helpful in training, quality control, and resolving questions as to why a particular action was taken.
- Link groups may also be printed or saved to an ASCII text file as needed.
- A status report can also be generated at any time during the resolution process. This report gives counts of the number of resolved and unresolved records by initial AutoMatch status. The counts are generated for each of the subprojects.

Summary

The new NASS record linkage system has been used operationally on a number of different matches. In 1997, responsibility for conducting the Census of Agriculture was transferred from the Bureau of the Census to NASS. Previously NASS and the Census Bureau each maintained a list of known agri-businesses and farm operators. With the transfer of the Census of Agriculture responsibility to NASS, the two lists were merged into one master list. The majority of the matches performed with the new record linkage system have involved merging various subsets records from the two lists. All matches were run in Headquarters and most of the possible matches were resolved in the State offices. Feedback was obtained from users on areas where the system could be improved.

The new NASS Record Linkage system is not yet fully developed. The back-end portions of the system used for on-line resolution and updating the master List Frame are complete with the exception of a few minor enhancements.

The front-end portion of the system is currently under development. This portion of the system will allow users to define the format of new source records and develop customized sets of match parameters. This portion of the system will be used by a limited number of users due to the technical knowledge and skill needed to develop match parameters. Extensive training will be needed to implement this portion of the system.