A Record Linkage Strategy to Match New Source Lists

Tom Pordugal, Kara Daniel, Stan Hoge, Bill Iwig

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

Key Words: Record Linkage, List Frame, SuperStan, SuperMatch

Introduction

The National Agricultural Statistics Service (NASS) maintains a List Frame of all known active, inactive and potential farm operators and agribusinesses. This list provides the population or sampling frame from which NASS gathers agricultural information through a system of survey programs. The List Frame is maintained at the state level by NASS's State Statistical Offices (SSOs). To keep the List Frame as complete and up-to-date as possible, NASS continually receives new data lists from a variety of sources. These new lists are matched against the existing SSO List Frame using probabilistic record linkage methodology. The NASS record linkage system uses Ascential Technologies' SuperStan and SuperMatch as the core software for its standardization and matching. A general overview of the record linkage matching techniques is described in the 1999 paper, "Record Linkage at NASS Using SuperMatch."

The NASS record linkage system has been used operationally for a multitude of different lists, including Farm Service Agency (FSA) programs, Farm Census, Federal, breeding associations, livestock and crop associations, farm bureau , seed growers, state veterinary services, and marketing associations. New source lists often contain different types and quality of information requiring different matching strategies. At a minimum, the lists contain name and address information. Different matching strategies, which include the blocking of variables within a pass, the matching variables used within a pass, and the determination of the number of passes, are required for different new lists.

The NASS record linkage system utilizes both front- and back-end applications. The back ends provide a tool whereby the results of a record linkage match can be clerically reviewed. The back-end portion of the system uses real-time database resolution and has capabilities to update the (master) SSO List Frame. The front-end portion of the system is developed in conjunction with the record linkage software features of SuperStan and SuperMatch. The front ends incorporate the reformatting and standardization of the new source file, or A file, and the master List Frame source, or B file, using SuperStan and then incorporate the matching parameters of SuperMatch.

The remainder of this paper discusses the features of the front ends and will follow the processing of a specific new source list from the initial raw data to final match output, which is then used to populate the back-end portion of the system for the SSO resolution review. An overview of the steps taken to match a new source list to the NASS List Frame is documented in the paper, including features of the record linkage software used to process the data, guidelines for developing matching strategies, and how NASS administrative data are used in the matching process. Specifically, the paper documents the steps applied to a new source list, namely, an FSA livestock compensation list.

Step 1: The New Source Request Form and Identifying the New Source Characteristics

Generally, each time that a new list source is to be processed through the record linkage system, a new list source request form is filled out by Headquarters or SSO personnel. If the data are obtained at the state level, then SSO personnel fill out the request form and transmit the form and the data to Headquarters. This form helps determine the characteristics of the new list. Key information includes both objective and subjective questions. The objective questions provide data that are hard coded into the reformatted file. These data relate to administrative information required by the master List Frame database. Subjective

information includes the quality of the list, special name and address rules for the matching process, the expected number of potential new farms added to the master List Frame, and special back-end administrative handling of the matched records and new records added to the master List Frame.

For this paper, the new source list is a list obtained at the national level from FSA and is a livestock compensation list. The format of the data is Excel. The records on the list have up to two names and addresses, a Social Security Number (SSN), an Employer Identification Number (EIN), a phone number, a county code, and cattle, dairy cattle, sheep and goat check data items from the livestock compensation program sign up. The list does not guarantee separate, independent farm producers for each record, since the producers may have an operating arrangement at the farm or operation level based on commodity shares. For example, landlords, partners, multiple partners, or spouses may have signed up with FSA on a share basis. So the list would contain records related to the same livestock. Moreover, the complexity of shares with the FSA livestock data makes any numeric matching comparisons on the check data to the master List Frame difficult. The check data will be populated to the back-end application for the reviewer to see but will not be applied in the matching processes.

Step 2: Reformatting the Data

New source lists do not have common formats or layouts, and all the lists do not have the same data fields. They generally contain person name and address information and often contain additional secondary operation names and addresses. The NASS record linkage system is designed to process two names and addresses, two phone numbers, an SSN, an EIN, a county id, and up to six commodity items, like crop and livestock check data. A generic 399-text file layout has been defined. New source lists first need to be reformatted to a specific field position and field length for the subsequent standardization and matching processes.

For the FSA livestock data, once the format and data fields have been determined, the file needs to be cleaned and reformatted as an ASCII (American Standard Code for Information Interchange) text file to the required 399-record length, which is the input file for standardization. SAS code is used to clean the data and reformat the data. Each piece of data is reformatted and the file is written out, following the required field positions and lengths. The data in the reformatted file will eventually be used to update matched records on the master SSO List Frame and add new records to the master SSO List Frame.

Likewise, the master SSO List Frame is extracted and reformatted to a generic 399-record length. NASS administrative data are included in the extract. Administrative data include farm active status codes; codes at the farm operation level that link other parties, like partners or owners; codes at the producer level that link other associated operations; cross-reference codes that link farm records in unique ways; and codes that flag records as agribusinesses. The fields in both files are renamed to those used in the record linkage system and the List Frame. Both reformatted files are now ready to be transformed into the standardized 967-record length files.

Step 3: The Standardization Process

The standardization processes of SuperStan involve the creation of a match key or dictionary. The match key consists of individual fields from each free-format element of the name and address. The match key for the NASS record linkage system consists of 34 name fields and 22 address fields. The match key also defines 4 name arrays that use 22 of the 34 name fields and 6 address arrays that use 10 of the 22 address fields. Since the NASS List Frame maintains both individual and operation names and addresses and the FSA new source list has this information, the overall standardization process includes two address and two name standardizations that run independently of each other. Wholly, there are four standardization processes, and the standardizer program must scan the free-format name and address information, recognize each element and move it to the appropriate field in the match key. For example, the given name of an individual should always be placed in the given name field of the match key, regardless of where it was recognized in the input record. In addition, each of the key words in the match key has standardized abbreviations. Generally, the standardization name and address processes each use their own match key dictionary to define the format of the match key, a classification table to define how to

interpret the many key words found on the input files, and a pattern table to define the rules for converting the input record to the match keys.

Standardization begins by separating all of the elements of the name or address into tokens. Each token is a word, a number, or an alphanumeric mixture separated by predefined separators. At the same time the tokens are formed, each token is classified by looking to see if it is in the classification table. If it is in the classification table, then it is given the class indicated by the table, else it is given another class type. The pattern file contains the rules by which the standardization is accomplished. The major power of SuperStan rests with its ability to process powerful and flexible rules.

NASS uses an in-house process to standardize elements of the place match key for the city, state and zip code. The in-house process uses its own place dictionary, which is periodically extracted from the NASS List Frame and conforms to the United States Post Office standards. Also, another process validates elements of the SSN, EIN and phone match key.

Because there are two name standardization processes, a final verification process is used to make sure that the person and operation names are in the proper fields of the match key. New source lists do not always place the names of a certain type in the same field. As the names are run through SuperStan, they are classified according to their operation type. There are three operation types, namely, corporation, individual operation, and partnership. The program looks at the SuperStan operation classifications and rearranges the information accordingly. If the information is switched on the new source list, then the standardized information is also switched. For example, if the operation name is missing on the new source and there is an operation or partnership in the person field, then the original data and standardized data in the operation name and whole name fields are switched. The name verification process is important because it helps improve the accuracy of the name blocking of SuperMatch.

For each of the lists, the final output contains a newly created match key. Each list now has individual parsed data fields, followed by the original 399-record length input record, resulting in the 967-record length file. A dictionary file defines the variable layout and how missing values are represented. For example, each field is defined appropriately for missing values, i.e., how the missing values are represented, like spaces or zeroes. The output from the SuperStan program can be used directly by the SuperMatch record linkage system for name matching, address matching, and new source list unduplication.

Step 4: The Blocking and Matching Phase

Knowing the information about the new source data provides insight into the kind of strategies needed for the matching process. SuperMatch provides a one-to-many file matching, based on individual records in file A matching one or many records in file B, and an unduplication, based on matching similar records within file A alone. The FSA list needs record unduplication since similar names, addresses, phone numbers, SSN and EIN information may be present. This unduplication of records will help identify and group records at the farm level when joined to the master List Frame records.

The SuperMatch software brings records together through a series of passes where different blocking variables, matching variables and thresholds, or cutoffs, are used. SuperMatch links records using the probabilistic record linkage techniques proposed by Fellegi and Sunter in 1969. It assigns a component weight to each of the variables (fields) being compared, according to the probability associated with the field. 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, which classify each case as a match (above the upper cutoff), non match (below the lower cutoff) or possible match (between the upper and lower cutoff). Throughout the matching process the master List Frame is treated as the reference B file. All new source records and List Frame records are included in each pass.

SuperMatch uses two phases for matching: the first phase partitions the data into subsets or blocks and the second phase associates a record on one file to a record on another file or on the same file (in the case of unduplicating a file). The first phase is referred to as the blocking phase and the second phase as the matching

phase. A pass is created for each block and match criteria.

The blocking phase limits the number of record pairs being examined, and increases the efficiency of the matching. This phase creates a block of records that have a potential of being associated with or linked to other records during the matching phase. All records having the same value in the blocking fields are eligible for comparison during the matching phase. For example, if NYSIIS of surname is a blocking field, all persons with surnames that fit the same NYSIIS algorithm are included in the block for the matching phase. (The NYSIIS code is a phonetic coding scheme used to reduce the effect of different spellings of the same name.) Any records with different NYSIIS surnames not included in the matching phase become residuals for this pass. With subsequent passes different blocking fields should be specified so that the residuals are eligible for the matching phase.

The matching phase is complex and the following matching parameters need to be defined: the comparison types and the associated m- and u- probabilities for each matching field, the weight overrides which assign replacement weights when the normal method of calculating the component weights are not appropriate, the cutoff weights which classify each case as a match, non match or possible match, and the assignment of actions on certain fields which force special classification treatment.

The comparison types are algorithms based on the type of data in the field, such as character data. Typical character data include zip codes, house numbers, box numbers, phone numbers, etc. The name and street address fields use a name uncertainty algorithm which is different than the usual character-to-character comparison. The algorithm assigns an adjusted weight based on differences between the two strings being compared. The algorithm tolerates phonetic misspellings, transpositions, and random insertion, deletion or replacement of characters but is dependent on a pre-assigned tolerance level for errors. Note that the full agreement or disagreement weights are applied to character comparison variables. Certain fields, like the name suffix fields, have clerical review actions assigned to them, where disagreement forces the record pairs being considered into automatic clerical reviews.

The m-probability is the probability that the field agrees given the record pair is a match. Namely, the m-probability is one minus the error rate of the field. That is, if a field in a sample of matched records disagrees ten percent of the time, then the m-probability for this field is 1 – 0.1 or 0.9. The u-probability is the probability that the field agrees given the record pair is unmatched, i.e., that the field agrees at random.

Fellegi and Sunter (1969) extended the concepts of m- and u- probabilities into a more rigorous mathematical treatment. Agreement and disagreement weights are used to measure the contribution of each field to the probability of making an accurate classification. Their definition of assigning weights takes into account the error probabilities for each field by using a log-likelihood ratio.

In general, the name and address fields are important matching fields. For most NASS applications these fields have pre-assigned m-probabilities of .90 or higher and are not revised during the matching process. The u-probabilities are also pre-assigned but are generally based on frequency counts. That is, the number of occurrences of that value of a field divided by the number of total occurrences of all values is equal to the u-probability. For variables that are uniformly distributed, like SSN, EIN and phone, the frequencies are not calculated and the pre-assigned u-probabilities are used.

Step 5: A Multiple Pass Design and the Effect of NASS Administrative Data

The determination of the number of blocking and matching passes depends on the information that the new source data contains and the quality of the new source data. A multiple pass design is used to increase the number of matched records, given errors in the blocking variables. Ideally, the possible matches from one pass are reviewed and resolved before the next pass is run. Because of the time and resources needed to review the possible matches between each pass, management of the back-end processes can become cumbersome. NASS conducted research to determine the consequences of performing one clerical review of the possible matches after all passes have been run, rather than a clerical review between each pass. This research found that the error rate of a combined review was not high enough to warrant the additional complexity of reviewing records between each pass. In the

NASS record linkage system, records are matched using the SuperMatch software, but the output files for all passes are combined, and only one clerical review is performed after all passes are run. See the matching techniques described in the 1999 paper, "Record Linkage at NASS Using SuperMatch".

Certain fields, like SSN, EIN, and local phone number, are considered key matching variables. NASS wants records with the same values for certain key variables like SSN, EIN and local phone to come together for the final review despite other information in the record. So, the normal method of calculating the component weights is not appropriate. To do this, a series of three passes is set up so that records with the same values for these variables come together. SSN, EIN and local phone matching variables are given weight overrides of 80. Since SSN, EIN and local phone are the blocking and matching variables for passes 1, 2 and 3, respectively, note that the composite weights of all the matched pairs in each of the passes is automatically scaled to 80. Specific name and address fields, like given name, surname, operation name, street number, box number and zip code are given weight overrides in these passes. These additional components are equally weighted, and each has weight override values of 3. The weight override value of 3 is nearly equivalent to an m-probability of .90. Also, the area code is given a weight override of 3 in the phone number pass. Agreement or disagreement of these additional items beyond the key matching variables is used to determine whether the records are classified as matches or possible matches. In effect, the passes are not really probabilistic.

Table 1 illustrates the design of the between and within list matching strategy of the FSA new source list. From Table 1 there are 16 passes defined using the between list matching. Note that passes 1, 2, 3, 9 and 10 bring records together at least as possible matches based on an exact match of SSN, EIN, local phone, or 911, rural route and PO Box delivery addresses. Passes 4 through 8 are used to refine true match pairs from the possible matches based on specific person name blocking. Passes 11 through 16 are used to refine true match pairs from possible matches based on specific operation blocking. These 11 name passes are also used to refine a possible match from a non match pair. The match counts and possible match counts are illustrated in the table. From the between list matching, there are 9,461 of 10,087 unique FSA records that at least possibly matched to the SSO List Frame. This leaves the remaining 626 records as potential non matches. Any duplication within 626 records, forces duplication and adds back to the possible match classification. It's the threshold level or cutoff in each pass, which classifies each pair as a match (above the upper cutoff), possible match (between the upper and lower cutoff) or non match (below the lower cutoff). Note that when the FSA records link to NASS records, none of the FSA records are excluded from subsequent passes.

From Table 1 there are an additional 8 passes defined using the within list matching or unduplication matching. Like the between list matching, the strategy here is to bring records together at least as possible duplicates based on an exact match of SSN, EIN, local phone or street address. Passes 1, 2, 3, 5, and 6 are used for this purpose. The other 3 passes refine the true duplicates from the possible duplicates and the possible duplicates from the non duplicates. The match duplicate counts and possible match duplicate counts are shown in the table. From the within list matching, there are 1,451 of 3,082 target (parent) records and the remaining 1,631 of 3,082 are at least possible duplicates to the 1,451 parent records. Note that any parent or duplicated FSA records can be included in the other passes. The remaining 7,005 of 10,087 FSA records were not considered duplicates.

There is specific NASS administrative data that forces matched records to be reviewed, whether the pairs are classified as matches or possible matches. The master List Frame maintains records at the farm operation level with links to other third parties, like partners, landlords or owners and the records have common operation identification numbers. The List Frame also maintains records at the producer level with links to other operations, and the records have common person identification numbers. The List Frame also maintains records based on link identification numbers across records. Specifically, when records in the A file are matched to records in the B file, and the B file records are duplicated to other B file records based on duplicate operation numbers, person numbers or link numbers from the master SSO List Frame, the records in the pair get classified as a possible match. Also, the master List Frame

maintains records based on active or inactive farm status. Any matches to inactive List Frame records force the link group to be reviewed. If the new source matches to ag businesses or out-of-state records on the List Frame, these also force a review.

Also shown in Table 1 are counts of forced review matches. These counts are originally considered match counts based on probabilistic methods, but due to the NASS administrative data attached to matched pairs, the records are considered suspicious matches and have been moved to the possible match counts requiring clerical review. The effect of NASS administrative data greatly impacts the review of matched records. The forced clerical match counts in the table are based on any of the following conditions in the NASS administrative data: the FSA records are matched to inactive records, records with person, operation or link identification numbers, records with an out-of-state address, or records with an ag business flag. Note that the entire SSO master List Frame currently contains about 92,000 of 157,000 (or 59%) records with this type of administrative data.

There is another step that creates the final link groups, which are then populated to the back-end application for SSO review. This is a complex SAS program that combines different sets of output files together to form the link groups. The first part of the program brings links with common FSA records or NASS records together. Generally, the program combines the different sets of output files together to form link groups for review. For example, if records A and B are linked as possible matches in one pass and the same two records are linked as matches in a second pass, the program classifies them as matches. However, if records A and C are linked as a possible match in a third pass, the program reclassifies records A, B and C as a possible match link group. See Table 2 for examples of combinations across multiple passes.

The second part of the program brings records together based on NASS administrative data. The program brings records together based on similar master List Frame person identification numbers, operation identification numbers and link identification numbers. Then, the program creates the actual link groups that are populated to the back-end application. The link groups in the back-end application are broken out by project, classification and review status. Table 3 shows the counts of link groups by classification and review status and project. SSO personnel are required to review and resolve all unresolved link groups.

Summary

The goals of maintaining a high coverage of farm operations on the NASS List Frame is difficult since operations in the United States do not have official registered names, and the farms on the NASS List Frame are activated or inactivated real time. When new list sources are matched against the List Frame, knowing the quality of the list and the limitations of the data help in the development of good matching strategies. Using probabilistic record methodology provides an excellent tool in bringing records together. However, since farm operating arrangements are complex and are maintained as complex entities on the List Frame, additional steps are taken to bring records together based on NASS administrative data from the master List Frame. NASS continually works on improving strategies to define the parameters used for its record linkage projects. As new ideas and suggestions surface to make the record linkage procedures more accurate and efficient, NASS constantly struggles to achieve a balance between reviewing the fewest possible match record groups and maintaining low duplication and match error rates.

References

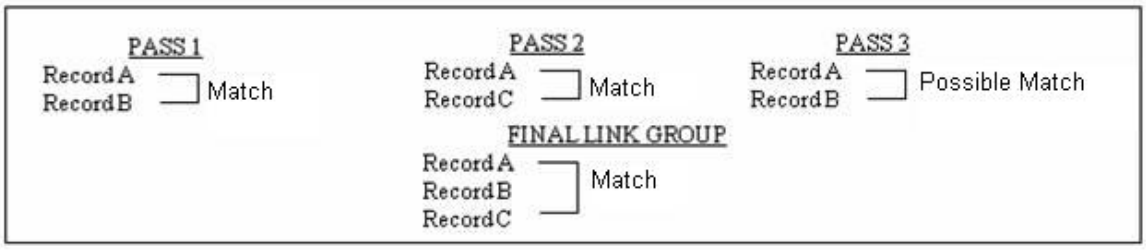
Fellegi, Ivan P. and Sunter, Alan B. (1969). *A Theory of Record Linkage*. Journal of the American Statistical Association. 64:1183-1210.

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

Table 1: A Matching Strategy of the FSA Source List with Match Counts for a Specific SSO
FSA Data (n=10,087) and SSO Data (n=155,867)

Pass	Blocking and (//) Matching Fields w/ Weight Overrides	Matches	Possible Matches and (+) Forced Review Matches
1	SSN // (SSN, Names & Addresses) w/ SSN: AW= 80 & Specific Name and Address Variables: AW= 3	5,745	2 + 1,552 = 1,554
2	EIN // (EIN, Names & Addresses) w/ EIN: AW= 80 & Specific Name and Address Variables: AW= 3	257	0 + 155 = 155
3	Local Phone // (Local Phone, Area Code, Names & Addresses) w/ Local Phone: AW= 80, Area Code: AW= 3 & Specific Name and Address Variables: AW= 3	4,274	563 + 1,201 = 1,764
4	(Given Name, Middle Name, NYSIIS of Surname and City) // (Names, Addresses)	1,247	137 + 337 = 474
5	(Given Name, NYSIIS of Surname and City) // (Names, Addresses)	3,969	1,014 + 1,052 = 2,066
6	(NYSIIS of Surname & City) // (Names, Addresses)	85	982 + 53 = 1,035
7	(NYSIIS of Surname & County) // (Names, Addresses)	70	7,813 + 22 = 7,835
8	(NYSIIS of Surname) // (Names, Addresses)	5	151 + 0 = 151
9	(NYSIIS of Street & City) // Addresses	0	7,527 + 2 = 7,529
10	(Box Number & City) // Addresses	0	1,672 + 2 = 1,674
11	((Operation_ NYSIIS of surname (A file) & Person_ NYSIIS of surname (B file)) & City) // (Names, Addresses)	98	71 + 64 = 135
12	Operation_Operation Name // (Names, Addresses)	4	3 + 1 = 4
13	(Person_Operation Name (A file) & Operation_Operation Name (B file)) // (Names, Addresses)	26	12 + 18 = 30
14	(Operation_Operation_Keyword1 & City) // (Names, Addresses)	11	2 + 6 = 8
15	((Person_Operation_Keyword1 (A file) & Operation_Operation Keyword1)& City) // (Names, Addresses)	208	72 + 108 = 180
16	Zip Code // (Names, Addresses)	62	157 + 39 = 196
1	SSN // (SSN, Names & Addresses) w/ SSN: AW= 80 & Specific Name and Address Variables: AW= 3	35	35
2	EIN // (EIN, Names & Addresses) w/ EIN: AW= 80 & Specific Name and Address Variables: AW= 3	6	10
3	Local Phone // (Local Phone, Area Code, Names & Addresses) w/ Local Phone: AW= 80, Area Code: AW= 3 & Specific Name and Address Variables: AW= 3	152	163
4	(NYSIIS of Surname & City) // (Names, Addresses)	585	669
5	(NYSIIS of Street & City) // Addresses	390	454
6	(Box Number & City) // Addresses	97	103
7	NYSIIS of Surname // (Names, Addresses)	240	252
8	Zip Code // (Names, Addresses)	529	601

Table 2: Combination Diagrams for Linked Records in Multiple Pass Runs



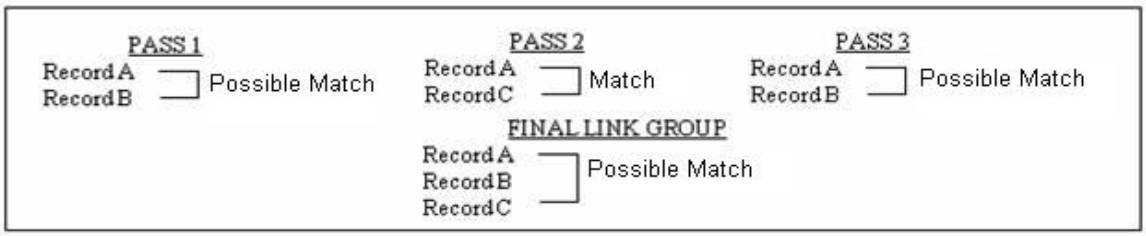


Table 3: Final Link Group Counts of the Matched FSA Source List for a Specific SSO
A 2-Way Table of Classification and Resolution Status By Link Group Type

		Description of Link Groups				
Classification	**Resolution Status**	Out of State Farms	Multiple Farms	Inactive Farms or Ag Businesses	Active Farms	**Totals**
Definite Matches	Resolved				3,982	3,982
	Unresolved	27		517		544
Possible Matches	Resolved					
	Unresolved	37	1,614	294	1,624	3,569
Non Matches	Resolved				520	520
	Unresolved	30				30
Link Group Counts		94	1,614	811	6,126	**8,645**