



United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research Division

STB Research Report
Number STB-96-02

October 1996

Record Linkage III: Experience Using **AUTOMATCH** in a State Office Setting

Kara Broadbent

RECORD LINKAGE III: EXPERIENCE USING AUTOMATCH IN A STATE OFFICE SETTING, by Kara Broadbent, Technology Research Section, Research Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250-2000, October 1996, Report No. STB-96-02.

ABSTRACT

The National Agricultural Statistics Service relies heavily upon a list sampling frame to select operators for agricultural surveys. Maintaining this frame is time-consuming and expensive because operations constantly change and new list sources are frequently introduced. The technique of identifying duplication within a file and matches between files is called record linkage. NASS currently uses a record linkage system called the Record Linkage Sub-System (RECLSS) to maintain the master list file and merge new sources with the master. In 1992, NASS decided to replace its current list frame database with a Sybase system called ELMO. This change stimulated research into alternative record linkage systems. A package named AUTOMATCH was evaluated as the basic component of the record linkage system for ELMO.

This paper presents a comparison of AUTOMATCH with RECLSS. The programs were evaluated by matching a list of potential new operations, obtained from the Ohio Producer Livestock Marketing Association, with the Ohio List Frame. The number of correct and incorrect decisions made by each program was recorded. Error rates were then calculated. These rates suggest that AUTOMATCH does indeed perform substantially better than RECLSS as it is currently used. However, some programming, which is already underway, needs to be completed in order to make the program user interfaces simpler to use, and automate the matching process for common files. With the addition of the Census of Agriculture to the NASS program, it is especially important that the final AUTOMATCH refinements are implemented.

KEY WORDS

Record Linkage; List sampling frame; Duplication; Software; AUTOMATCH.

The views expressed herein are not necessarily those of NASS or USDA. This report was prepared for limited distribution to the research community outside the U.S. Department of Agriculture.

ACKNOWLEDGMENTS

Much of the work for this project was conducted while the author was a member of the Field Research Unit, which was located within the Ohio State Statistical Office. The author would like to thank members of the Ohio SSO for their guidance throughout the project. A special thanks to Jacqui Thomas for her help in evaluating the mainframe and AUTOMATCH output. Thanks to Charles Day for resourceful insights and technical advice. Thanks to the management (Ron Bosecker, George Hanuschak, Roberta Pense, and Lee Brown) for their support of this project.

TABLE OF CONTENTS

SUMMARY	iii
INTRODUCTION	1
BACKGROUND	2
METHODS	6
RESULTS	9
North Carolina Match	12
Ease of Use	13
Software Problems	14
CONCLUSIONS	15
RECOMMENDATIONS	16
REFERENCES	17

SUMMARY

The National Agricultural Statistics Service (NASS) gathers information concerning various aspects of agriculture through a system of surveys. NASS uses a multiple frame sampling design to conduct some of its surveys. Samples are drawn from both a list and area frame for multiple frame samples. The area frame consists of the total land area in the United States, divided into sampling units. The list frame is a register of known farm operators and agribusinesses. Maintaining a current, accurate list frame is a time-consuming and expensive process. As new list sources are obtained, they must be unduplicated and matched against the master list frame file. The technique of identifying duplication within a file and matches between files is called record linkage.

Because an accurate list frame is so efficient for sampling, NASS devotes a considerable amount of time and resources to list maintenance. NASS currently uses a computerized record linkage system called the Record Linkage Sub-System (RECLSS). It is based on record linkage theory proposed by Ivan Fellegi and Alan Sunter in a 1969 *Journal of the American Statistical Association (JASA)* paper. In 1992, NASS decided to replace its list frame database, the Real Time Mail Maintenance System (RTMMS) with the new Enhanced List Maintenance Operations (ELMO) database. This change created an opportunity to evaluate new record linkage systems. Charles Day studied the available record linkage software packages and recommended that NASS adopt a commercial program named AUTOMATCH as the core component of the new system. Like RECLSS, AUTOMATCH uses the Fellegi-Sunter record linkage technology to provide a statistically justifiable methodology for matching.

The purpose of this study was to compare AUTOMATCH with RECLSS, both in terms of accuracy and ease of use. The two programs were compared by linking a list of potential new operations, obtained from the Ohio Producer Livestock Marketing Association (PLMA), with the existing operations on the Ohio List Frame. Counts were made of the number of correct and incorrect decisions made by each program. Finally, these results were used to create a master data file. This file associated each PLMA record with both corresponding list frame ID'S and duplicate PLMA ID'S. It will be used for future record linkage research in enhancing parameter specification files.

Results from this study suggest that AUTOMATCH does indeed perform substantially better than the RECLSS system as it is used today. The number of false matches and false nonmatches made by AUTOMATCH were both lower than the RECLSS system. Despite the fact that the error rates for AUTOMATCH were lower than those of RECLSS, AUTOMATCH has several areas where improvements are needed before it is made fully operational. Good front and back end graphical user interfaces need to be created to make the program more user friendly. Developing default matching specifications, thresholds, and parameter values for commonly used files, would also simplify the matching process. With the addition of the Census of Agriculture to the NASS program, it is especially important that the final AUTOMATCH refinements are implemented.

INTRODUCTION

The National Agricultural Statistics Service (NASS) gathers information concerning various aspects of agriculture through a system of surveys. Farm operators and agribusinesses are regularly sampled in order to obtain statistical estimates for commodities. The quality of these estimates is partially dependent upon the quality of the sampling frames used.

NASS uses a multiple frame sampling design to conduct some of its surveys. Samples are drawn from both a list and area frame for multiple frame samples. Using this combination allows samples to be drawn which maintain the advantages of both frames. The area frame consists of the total land area in the United States, divided into sampling units. This type of sampling frame provides complete coverage of the geographical area of interest, despite operation changes. The area frame supplements the list frame by providing a measure of incompleteness.

The list frame is a register that includes names, addresses, and control data of all known farm operators and agribusinesses. Having control data allows stratified samples to be drawn and provides an efficient method of sampling. However, the list frame does not completely represent all farm operations in the population. Maintaining a current, accurate, and relatively complete list frame is a time-consuming and expensive process. Farm operations are constantly changing and new list sources are often incomplete.

Because an accurate list frame provides such an efficient sampling frame, NASS devotes a considerable amount of time and resources to list maintenance. The frame must be continu-

ally updated with new farm operators and agribusinesses, which are obtained from a variety of sources. Names and addresses on these new source lists must be standardized so that they can be machine-compared to identify duplication. Duplication both within the new source lists and between the new lists and master file must be identified to ensure that each record added to the master file represents a unique farm operation. The technique of identifying duplication within a file and matches between files is called record linkage.

NASS currently uses a computerized record linkage system it developed in the late 1970's. This set of programs is called the Record Linkage Sub-System (RECLSS). RECLSS is used both to maintain the master list frame data file and merge new source files with the master. It is located on a main-frame computer system used by NASS on a contract basis. The statistical foundation of this system is based on the record linkage theory proposed by Ivan Fellegi and Alan Sunter in a 1969 *Journal of the American Statistical Association (JASA)* paper [1]. RECLSS provides a thorough standardization and matching system, but it is time consuming and expensive to operate. It is complicated and requires experienced users to achieve accurate results. Because of its complexity and lack of current documentation, users often perceive RECLSS as a black box.

In 1992, NASS decided to replace its main-frame list frame database, the Real Time Mail Maintenance System (RTMMS), with a new Sybase client-server database, called the Enhanced List Maintenance Operations (ELMO). This change created an opportunity to explore new record linkage solutions. Charles Day evaluated the available record

linkage software packages and recommended that NASS adopt a commercial program named AUTOMATCH as the core component of the record linkage system for ELMO [2]. AUTOMATCH was developed by Matthew Jaro of MatchWare Technologies Inc. Like the mainframe system, AUTOMATCH also uses the Fellegi-Sunter record linkage technology as a foundation to provide a statistically justifiable methodology for matching. In addition to its record linkage system, AUTOMATCH also includes a flexible name and address standardization package named AUTOSTAN. AUTOMATCH, like RECLSS, is also a time consuming and complex system to operate; however, it will potentially allow the user more flexibility than RECLSS as it is currently used.

The purpose of this study was to compare the operation and performance of AUTOMATCH, as it would be used, with RECLSS, as it is currently used, in a State Statistical Office (SSO) setting. The two systems were compared in terms of accuracy and ease of use. The research was conducted in the Ohio SSO, which was the location of the Field Research Unit. Performing the test in Ohio rather than Headquarters allowed us to evaluate the possibility of operating AUTOMATCH in a State Office. It also gave us the opportunity to assess the amount of time it took for new users to understand the background information necessary to run the software. Conducting the research in Ohio also gave the additional benefit of State Office support staff and resources to resolve any questions concerning potential matches.

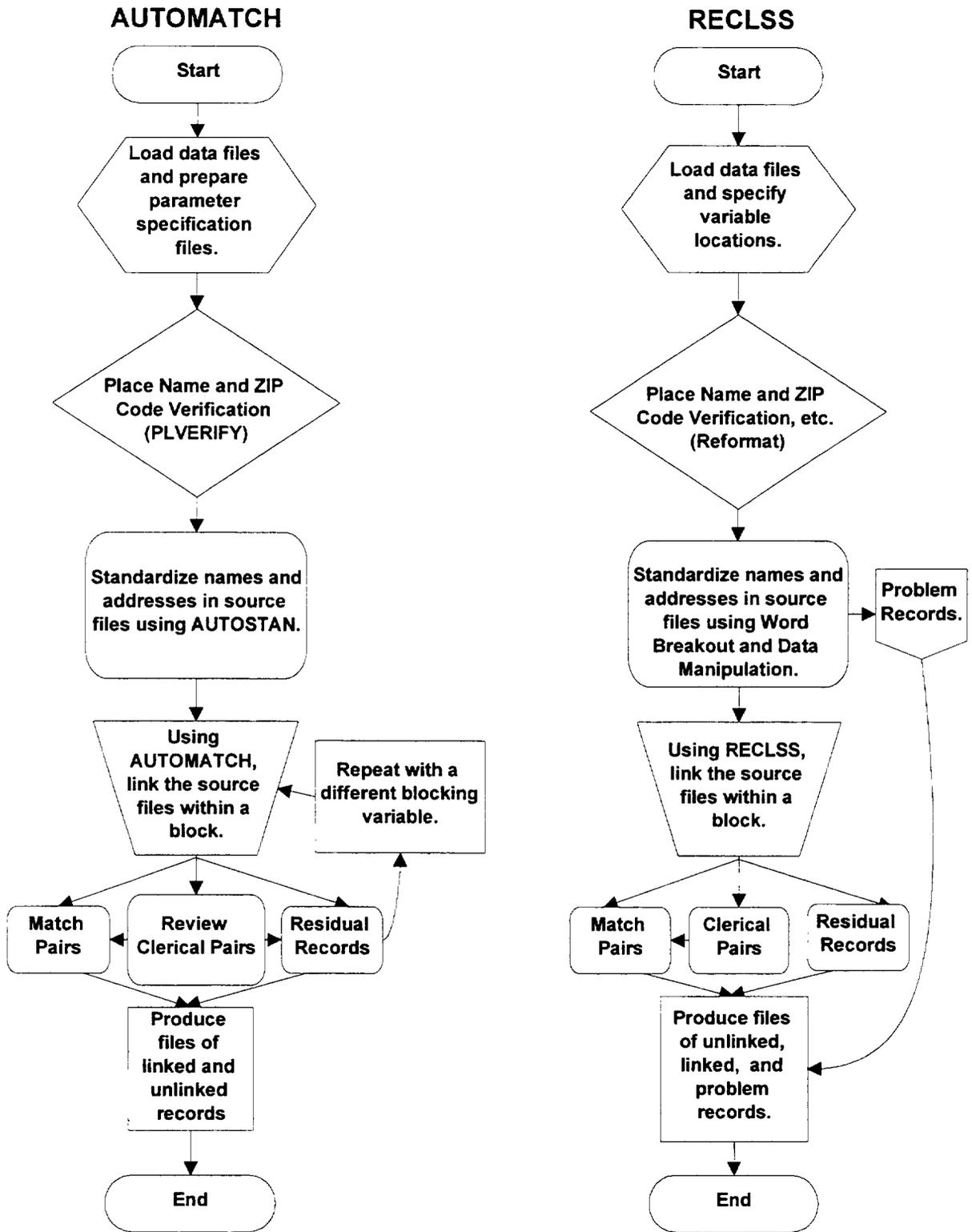
AUTOMATCH and RECLSS were compared by linking a list of potential new farm operations, obtained from the Ohio Producer

Livestock Marketing Association (PLMA), with the existing operations on the Ohio List Frame. These lists were standardized and matching was performed between and within the two lists by both the RECLSS and AUTOMATCH systems. Counts were made of the number of correct and incorrect decisions made by each program. Finally, these results were used to create a master data file. This file associated each PLMA record with both corresponding list frame identification numbers (LSF ID'S) and duplicate PLMA ID'S. It will be used for future record linkage research.

BACKGROUND

Matching was performed both within and between the PLMA and Ohio list files using RECLSS and a DOS version of AUTOMATCH. The following is a general description of the two processes. More detail is given to the AUTOMATCH process because it is being considered as a possible replacement for the RECLSS system. Diagram 1, on the following page, gives an overview of flow of AUTOMATCH and RECLSS as they were used in this comparison. Further information regarding RECLSS can be found in the list frame resolution manual [4]. Information regarding AUTOMATCH can be found in the documentation manuals [5,6] supplied with the software and in two research papers written by Charles Day. These reports contain background information on record linkage and NASS. The first entitled, *Record Linkage I: Evaluation of Commercially Available Record Linkage Software for Use in NASS* [2], discusses the reasons AUTOMATCH was selected by NASS. A subsequent paper, *Record Linkage II, Experience Using AUTOMATCH for Record Linkage in NASS*

Diagram 1.--Overview of AUTOMATCH and RECLSS Processing Used in This Study.



[3], summarizes NASS's record linkage experiences with AUTOMATCH. This paper also provides a general description of how AUTOMATCH operates. **Readers not familiar with NASS or record linkage should read those reports first.**

Overall, AUTOMATCH and RECLSS perform similar functions. Both systems standardize the input files, perform matching using the widely accepted Fellegi-Sunter record linkage theory, and generate files of output records. Records are designated as matches, nonmatches (residuals in AUTOMATCH terminology), or clerical review cases. Clerical record pairs are possible matches that must be reviewed by hand before determining their final status.

As a first step in standardization, place names and ZIP Codes are verified. Each ZIP Code is checked to be certain that it falls within the allowable range for its associated city. Corresponding latitudes and longitudes for the place names are then assigned to each record. This process provides more complete and accurate records for matching. The Reformat program in RECLSS performs the place name verification process. AUTOMATCH does not have a place name verification procedure. An in-house C program, named PLVERIFY, was written to perform essentially the same function as RECLSS's Reformat procedure.

Once the place names and ZIP Codes are verified, the records are ready for standardization. This process converts records from a free format to one with distinct fields for each element. This involves breaking the names and addresses into basic components such as title, given name, middle name, last name, house number, and street name. It also

converts nicknames and other abbreviations into a standard format. Standardization is a critical step in the unduplication process. A well-standardized list makes the matching process much more accurate. The Word Breakout and Data Manipulation programs perform the RECLSS standardization routines. For AUTOMATCH, standardization is done by AUTOSTAN, AUTOMATCH's companion program.

Although the functions of the RECLSS and AUTOMATCH standardization routines are essentially the same, the specific algorithms each performs are different. Both use similar classification tables to categorize the individual name and address components. However, the pattern tables used to recognize the record formats and generate the standardized fields are different. The way the two systems treat records that do not standardize is also different. If RECLSS cannot standardize any part of a particular record, it is written to an output file, as a problem record. It is not passed to the matching routines. With AUTOSTAN all records are written to the standardized output file, even if a portion of the record cannot be standardized. This allows all records to be passed from AUTOSTAN to AUTOMATCH and have the opportunity to be matched. The impact of this difference varies by list source and State. In many cases, the impact is minimal. However, in one recent linkage run where a new source was matched against a Midwestern State, 2,320 names could not be standardized by RECLSS. When these names were later run through AUTOSTAN, all but 116 names were standardized.

After the lists are standardized, the matching routines are performed by comparing the individual standardized components common

to both records. This type of comparison allows records to be matched regardless of the initial order of the variables or missing values within a record. Weights are assigned to each component based on probabilities. If two components are the same, a positive weight is assigned. If they are different, a negative weight is assigned. The individual weights are then summed to obtain an overall weight for the comparison pair. This composite weight is compared with two threshold values. If it is larger than the upper threshold, the pair is considered a match. If it is smaller than the lower value, the pair is considered a nonmatch. If it falls between the two, the pair is considered a possible match or "clerical."

The number of comparisons when two files are linked against each other is very large. Examining every possible match pair is not feasible. To make the task of matching more manageable records are assigned to blocks. A block is a mutually-exclusive subset of records which have the same value for a "blocking variable." Only records within the same block are compared with one another. All records in different blocks are considered nonmatches. This substantially reduces the number of comparisons made. An example of a possible blocking variable is ZIP Code. If two records had different ZIP Codes, they would be considered nonmatches. The only records that would be compared for matching are those which have the same ZIP Codes. Care must be taken to select blocking variables that are reliable. An incorrectly reported blocking value may cause a match to be overlooked.

AUTOMATCH uses multiple passes with different blocking variables in its matching routines. This decreases the number of com-

parisons made while still reducing the effect of incorrectly reported data. If a value for a particular blocking variable is reported incorrectly, the record would still have the potential of matching on later passes when different blocking variables are used. Between each pass of AUTOMATCH, a clerical review is conducted. This allows records classified as nonmatches after the clerical review of early passes to continue to be compared in later matches. The clerical review process is critical if optimal results are to be obtained. AUTOMATCH has an interactive computer assisted clerical review process; however, it has some serious limitations. Development of a independent clerical review process with a good graphical user interface will overcome these limitations and make this process feasible for use in practice.

Matching using RECLSS is similar to that of AUTOMATCH for individual records. RECLSS creates blocks and compares records within the same block for possible matches. It then uses the idea of Match Groups to identify linkage across blocks. A match group is a set of linked records from different blocks that are brought together because they have identical values for address (when a box or house number is present), EIN, SSN, telephone number, or link cross-reference number. When matching farm operations and partnerships with RECLSS, a nonprobabilistic matching routine is used. Manual review of clerical pairs is time-consuming since it is a batch process involving transaction coding. Consequently, it is often omitted in practice with the RECLSS system. Eliminating the clerical review process prevents matches across blocks from being detected. It also assigns all records that were classified as clerical review cases to be considered matches. This does

not allow RECLSS to perform to its full capability.

One distinct advantage of AUTOMATCH over RECLSS is that with AUTOMATCH the number of different matching schemes that can be created is not limited. With RECLSS matches can only be performed based on fifteen predetermined matching fields. AUTOMATCH allows matches on these fifteen fields in addition to any others which may be desired. The selection of blocking variables and matching parameters is also very flexible with AUTOMATCH.

METHODS

Commonly when record linkage programs are evaluated, small data files with known results are used. By matching the entire Ohio list frame against the PLMA list, a more extensive test was performed. Both data files were relatively large and the true results were unknown.

A 1994 extract of all active and inactive list frame records from the Ohio Real Time Mail Maintenance System (RTMMS) Name and Address master file was used for the study. There were 117,246 list frame records on this extract. Of these records, 63,285 were active (subject to sampling). The 1994 Ohio Producer Livestock Marketing Association register was also used. This is a list of all Ohio PLMA members given to the State Office every other year. Members of the association are farm operators who pay livestock marketing fees in the state of Ohio. There were 24,843 names on the PLMA list. The two files contained the following common identifiers:

- 1) primary name,

- 2) secondary name,
- 3) street address,
- 4) place name,
- 5) ZIP Code.

Before conducting the match of the PLMA and Ohio List Frame files, the list frame file was unduplicated using AUTOMATCH. It was assumed that there would be little internal duplication on the list. Of the 63,285 active list frame records, only 64 duplicates were found. The List Frame is examined for duplication using RECLSS as part of its regular maintenance process. Consequently, we did not feel it was necessary to unduplicate the list using both RECLSS and AUTOMATCH. As the output from the Ohio/PLMA match using RECLSS was generated, any possible list frame duplicates that may have been missed by AUTOMATCH were printed. No additional duplicates were found.

The List Frame Section in Headquarters performed the match between the PLMA file and Ohio List Frame using RECLSS. The only requirements of the State Office were to load the two files on the mainframe system and specify where the variables were located. Threshold parameters and other linkage options were carried forward from previous RECLSS linkages.

The AUTOMATCH test was run on a DELL 486-33 MHz personal computer with 16 Mb of memory and a 1 Gb hard drive. The 1 Gb hard drive was added because the original drive was not large enough to complete the matching process. Running AUTOMATCH required many more preparatory steps than RECLSS. Parameter specification files had to be generated defining variable locations, probabilities, thresholds and matching rou-

Figure 1. --AUTOMATCH Matching Scheme for Ohio List Frame, PLMA Match

<u>Pass</u>	<u>Blocking Variables</u>	<u>Linking Variables</u>
1	ZIP Code	Operation Name
2	ZIP Code	Primary and Secondary Name, Street Address, Place Name
3	Soundex Code of Primary Surname	Primary and Secondary Name, Street Address, Place Name

tines. The specification files are fairly complicated; however, very few changes were required once they were initially created. Once the specification files were generated, each pass was processed using a DOS batch job. Because it took so long to run a single pass, the process was usually run overnight. A single pass took approximately six hours to process. This time is substantially reduced when a pentium computer is used, and even lower on the UNIX system.

The matching routine shown in Figure 1 was used to run the AUTOMATCH test. It is a fairly simple process due to the fact that the PLMA list had a small number of variables. During the first two passes, records were blocked on ZIP Code. Operation name was used as the matching variable in the first pass. In the second pass, primary and secondary names, and address were used. Residual (nonmatch) records from the first two passes were blocked on Soundex Code of the primary surname for the third pass. They were again matched on primary and secondary names, and address. The Soundex Code is a phonetic coding system that reduces the effect typographical variations have on comparisons. RECLSS uses the NYSIIS coding system rather than the Soundex Code. When

this study was conducted, AUTOMATCH did not have the capability of generating NYSIIS codes. This function was added in a later release of AUTOMATCH.

RECLSS used a matching strategy similar to AUTOMATCH's pass three for individual names. It used a NYSIIS Code of the primary surname (similar to the Soundex Code) as the blocking variable and matched on names and addresses. It then used Match Groups to link records with different blocking variables. In order to compare the results of the two programs as we believe they actually are or would be used, **a clerical review of the RECLSS output was not performed.** Because no clerical review of the Match Group links was done, duplication across blocks was not detected. All records that were classified as clerical pairs during the RECLSS process were considered matches in the final output.

A clerical review was conducted after processing each pass of AUTOMATCH. At the time of this study, we believed this would be the way the software would be used operationally. The AUTOMATCH software contains its own clerical review processes called Clview. The screen format for the variables

during the clerical review process can be specified using one of AUTOMATCH's parameter files. Clerical reviews for this research were not done initially using the Clview process. Clview only allows one user at a time to review the records. It does not have the capability for the user to add comments or access the list frame. Once a record has been reviewed, reviewing or changing the status of that record again is not possible. Clview also does not allow the user to access a particular record by ID for resolution. To avoid these limitations, the clerical records were printed and reviewed on paper.

The paper copies of clerical records were also formatted using AUTOMATCH'S Repgen procedure. Repgen is a tool within AUTOMATCH which generates reports which can be viewed either electronically or as printed output. One drawback of Repgen is that when it formats the output there is nothing to distinguish one group of clerical records from another. This problem was overcome by writing a SAS/ IML program that reformatted the output by inserting a blank line between each clerical group. Examining the clerical records on paper allowed the review to be done more easily. It also provided a written history which could be reviewed or later changed.

As the clerical records were reviewed, each record on the printout was color-coded as a match or nonmatch. State Office resources were used to arrive at a final match status for non-obvious record pairs. After resolving all clerical records for a particular pass, AUTOMATCH'S clerical review program was run and the final decisions were interactively entered from the paper copy to the computer. This allowed clerical records that were nonmatches in early passes still to be

considered as potential matches in later passes. For all three passes, there were a total of 1,692 clerical records. It took approximately 120 staff hours to complete the clerical review.

The time that was spent reviewing clerical pairs is not indicative of that which would be required of a State Office when AUTOMATCH is used in practice. The review conducted in this study was much more thorough than that which would be performed operationally. The time also includes time spent transferring the final decisions from the printout to the computer.

After completely processing the data files through both RECLSS and AUTOMATCH, final output files with all matches and non-matches from the two programs were printed and compared. It took approximately six months to complete this process. Each PLMA record was examined to see if it was classified the same way by both systems. This involved locating each PLMA record on the output files from both systems. The electronic versions of each file were used to search for records. Once the same record was found on both systems, the statuses were compared. This process would not be required in practice. It was only conducted in order to measure the accuracy of the two systems.

The RECLSS system generated four output files which contained matches, and one file of nonmatches. In addition to the match and nonmatch files, RECLSS also generated several files that contained records which did not standardize correctly. Because there were multiple files generated for each type of record, determining which file a particular record was located in was difficult. The files

were hard to work with because the format was awkward. A potential match pair consisted of fourteen lines of output. We reformatted this output so that a pair only took up three lines (one line for the list frame record, one line for the PLMA record and a blank line to separate pairs). This output was printed on wide paper to use for comparison with the AUTOMATCH results.

It was assumed that when RECLSS and AUTOMATCH reached the same conclusions concerning a particular PLMA record that a correct decision was made. Examining each individual case would have required more personnel resources than were available. It was also assumed that if a PLMA record was matched to a list frame record that was link-cross-referenced to the true matching list frame record, a correct match was made. The cases we were primarily concerned with were those where the two systems reached different conclusions.

If AUTOMATCH and RECLSS reached a different conclusion concerning a particular record, it was examined further by an enumerator in the Ohio State Statistical Office to determine which program was correct. Often the list frame contained adequate information to make a decision. If not, the name was looked up in the phone book. At times this led to a decision, or gave a number to reach the operator for individual questions. The PLMA file did not contain phone numbers. This made the phone book and directory assistance our only sources for phone numbers when follow-up calls were necessary. Ohio has a relatively large population of Amish farmers who traditionally do not have phones. They typically have similar biblical first and last names and non-unique addresses (Rural Route 1 with no house

number). For these and other people without listed phone numbers, we used our best judgement when making a final decision.

Once all the differences in the AUTOMATCH and RECLSS output files were resolved, the final results were used to create a master PLMA file which linked each record to its associated matches and duplicates. This file contained the original information for each PLMA record, a PLMA ID number, the matching list frame ID'S for the record (if any), and any ID'S of other PLMA records which it duplicated. This file, along with the 1994 Ohio list frame extract, will be used in future research to develop default parameter specification files.

After completing the comparison of RECLSS and AUTOMATCH, a test was conducted to learn the consequences of conducting only one clerical review after all passes had been processed rather than separate reviews between each pass. Ideally a clerical review would be conducted after each pass. This allows clerical records which are made residuals in early passes to be included as potential matches in later passes. The PLMA and list frame files were matched twice; once with clerical reviews being done between passes and once with a single review after all passes were completed. The results were then compared to determine the differences in the two procedures.

RESULTS

When comparing the results from RECLSS and AUTOMATCH, we considered six outcomes. (The two cases where RECLSS and AUTOMATCH both said a record was a match, or nonmatch were not considered. An error could have been made by either

Figure 2.--Possible Record Linkage Outcomes

	Records represent the same unit in the population.	Records represent different units in the population
Records are linked by the record linkage process.	True Match	False Match
Records are not linked by the record linkage process.	False Nonmatch	True Nonmatch

program, however, the probability that both programs reached the correct decision was high. Resources were not available to examine each of these cases.) The six cases were:

- 1) AUTOMATCH said the record was a nonmatch, RECLSS said it was a match; in reality it was a match.
- 2) AUTOMATCH said the record was a nonmatch, RECLSS said it was a nonmatch; in reality it was a nonmatch.
- 3) AUTOMATCH said the record was a nonmatch, RECLSS said it was a match; in reality it was a nonmatch.
- 4) AUTOMATCH said the record was a match, RECLSS said it was a match; in reality it was a match.
- 5) AUTOMATCH said the record was a match, RECLSS said it was a nonmatch; in reality it was a nonmatch.
- 6) AUTOMATCH said the record was a match, RECLSS said it was a nonmatch; in reality it was a match.

In four of the above cases errors were made by one of the programs. Cases one and five were errors made by AUTOMATCH, and cases three and six were errors made by RECLSS. A false match is when a record pair is classified by the software as a match when in reality it is a nonmatch. On the other hand, a false nonmatch is when a record pair is classified by the software as a nonmatch when in reality it is a match. Cases three and five were considered false matches, where cases one and six were considered false nonmatches. Figure 2 displays the four possible outcomes when a program made a decision concerning a pair of records.

The number of correct and incorrect decisions made by AUTOMATCH and RECLSS was determined, and is displayed in Figure 3. In every case, the number of correct decisions made by AUTOMATCH was larger than the number made with RECLSS. Overall, AUTOMATCH made 24,207 correct decisions, while RECLSS made 22,455 correct decisions. The actual total number of matches can be found by adding the number of correct matches and incorrect nonmatches made by either program (16,234 + 1,533 or 17,192 + 575). Similarly, the actual total number of nonmatches can be found by adding the number of correct nonmatches and incorrect matches made by either program.

Figure 3.--Ohio/PLMA Match Counts--RECLSS vs. AUTOMATCH

	AUTOMATCH	RECLSS	AUTOMATCH	RECLSS
Linked (Match)	(True Match)		(False Match)	
	17,192	16,234	61	855
Not Linked (Nonmatch)	(False Nonmatch)		(True Nonmatch)	
	575	1,533	7,015	6,221

The counts displayed in Figure 3 were used to determine error rates for each program. The error rates were calculated by dividing the number of incorrect decisions in a category by the sum of the number of correct and incorrect decisions in the same category. For example, the false match rate for RECLSS would be the number of incorrect matches made by RECLSS divided by the total number records RECLSS classified as matches. This is equal to $855 / (855 + 16,234)$, or 5 percent. The false match and false nonmatch rates are shown in Figure 4.

The PLMA list had a large amount of internal duplication. Besides having the same individuals listed multiple times, there were many cases where members of the same family were all listed as separate operators. As we studied the list, we found that the name on the PLMA list was not necessarily the farm operator, but the person who paid

the marketing fee. This was often a family member such as a child, parent, or spouse. Determining whether names fell in this category or if they were really different operators was difficult. When it was verified that there were two different operators at the same address, the record status on the list frame was changed to 67. This is an active status code (subject to sampling) which indicates that there are two or more separate operators at the same address. Using this record status will make future resolution easier.

The final number of duplicates found by each program was not determined. Only an overall number of duplicates was recorded. It was assumed that the duplication error rates for RECLSS and AUTOMATCH would be similar to their matching error rates. The procedures for unduplicating a file and matching two files are essentially the same.

Figure 4.--Ohio/PLMA Match Error Rates--RECLSS vs. AUTOMATCH

	False Match	False Non-match
RECLSS	5.0%	19.8%
AUTOMATCH	0.4%	7.6%

Figure 5.--Final PLMA Counts

	Match with List Frame	Nonmatch with List Frame	Total
PLMA Operations	15,755	5,746	21,501
Duplicate PLMA Operations	2,012	1,330	3,342
Total	17,767	7,076	24,843

Counting the number of clerical records found by each program would have required more staff hours than were available.

After resolving each case where RECLSS and AUTOMATCH reached different conclusions, a final status was determined for each record. The final counts are displayed in Figure 5. Of the 24,843 PLMA records, 17,767 operations were already present on the list frame and 7,076 did not match a list frame ID. There were a total of 3,342 duplicates on the PLMA file. Of these duplicates, 1,330 did not match with a list frame record and 2,012 did match a record on the list frame. This left 15,755 operations that matched with a list frame ID and 5,746 operations that did not match a list frame ID.

The new farm operations that did not match with a record on the list frame were added to the list frame with an inactive status code (not subject to sampling). Criteria letters

were sent to these operations, and once additional information is obtained the record statuses will be changed to active codes.

North Carolina Match

A similar comparison of AUTOMATCH and RECLSS was done by Charles Day using the North Carolina list frame and a new source file. This study is found in his report entitled *RECORD Linkage II, Experience Using AUTOMATCH for Record Linkage in NASS* [3]. The North Carolina list frame contained 77,000 records and there were 7,404 records on the new source file. The two files were matched using both AUTOMATCH and RECLSS. The results were compared in a manner which was similar to that which was used for this study. Figure 6 displays the error rates obtained in the North Carolina study. Considering differences in the review process and source files, the results from the Ohio match seem reasonable when compared with those obtained in the North Carolina match.

Figure 6.--North Carolina/New List Match Error Rates--RECLSS vs. AUTOMATCH

	False Match	False Nonmatch
RECLSS	4.6%	8.7%
AUTOMATCH	1.1%	4.9%

The primary difference in the two studies was that the North Carolina study was done in Headquarters. State resources were not available for use during the clerical review process or when the results from the two programs were compared. Decisions for clerical pairs were made based only on information from the list frame extract file and information from the new source file. If AUTOMATCH and RECLSS arrived at a different conclusion regarding a particular record and it was not clear which was correct, it was assumed that RECLSS made the correct decision.

Another difference in the two studies was that the North Carolina list frame extract only contained active records, while the Ohio list frame contained both active and inactive records. Quite a few PLMA records were linked to inactive Ohio list frame records. Many of these records were recently added and inactive because they did not have adequate control data. Others were records coded as out of business, retired, or deceased. As these cases were investigated, it was found that at times children or spouses would pay the fees in the operator's name even though the operator was not farming. However, in other cases, it was found that the operation had resumed business, or was incorrectly coded out of business, and the record status needed to be changed.

A third difference in the two studies was that the new source files did not contain the same number of variables. The new North Carolina list source only had one name field, where the PLMA file had both primary and secondary operation names. The North Carolina new list source did have one important additional variable which was phone number.

Phone number is often very helpful in discriminating between possible matches. Not having phone number made many of the Ohio possible matches seem even more ambiguous. This made the clerical review process even more important. As the number of discriminating variables decreases the importance of a clerical review increases.

Ease of Use

Charles Day traveled to the Ohio SSO and provided three days of training on the fundamental components of Record Linkage. This included the Fellegi-Sunter theory, the basic steps to run AUTOMATCH, and a few example matches. The training was very helpful. Understanding the matching process would have been very difficult without some initial guidance.

The documentation manuals supplied with the software were helpful in providing basic AUTOMATCH background information. They had a good explanation of the components and options for each to the specification files. The manuals had examples of the parameter files; however, the examples were not consistently from the same matching routine. The manuals did not contain documentation on all the possible features of AUTOMATCH. After studying the AUTOMATCH documentation we wrote our own "mini manual" to provide a concise reference of the general steps needed to run any match. This manual gave examples of each of the command lines and parameter files required for a single match and unduplication.

The matching specification files were developed and processed by one person. Because of the large amount of time required for AUTOMATCH to complete a pass, the

matches were either run overnight or over the weekend. This allowed the records to be ready for clerical review the next working day. An office enumerator performed the majority of the clerical reviews. Because a thorough review was done of each record this was a time-consuming process. To overcome the limitations of the clerical review and report generation modules, the clerical review was done using printed output. A well-designed clerical review package, which can access the list frame, would speed up this process.

Conducting a single clerical review after all passes have been processed would be simpler than conducting clerical reviews after each pass. When comparing the results of conducting a single clerical review versus three separate between pass reviews in this study it appears that one final clerical review would be sufficient. There were 116 additional records which were classified as matches when the clerical review was conducted between each pass. This is less than one percent of the total number of matches. The benefits of having a simpler clerical review process outweigh the disadvantage of missing a small number of cases.

Once we performed the first match, modifying the parameter files for subsequent runs was fairly easy. The same matching strategies could be used for all matches because the files followed a general pattern for each match. Rewriting the parameter files did not require a large amount of time. With the exception of occasionally miscalculating the location of a particular variable, we rarely had any problems running AUTOMATCH. We did have one small problem with the format of the PLMA input file. AUTOMATCH requires each of the input

lines to be of a fixed length. A few lines in the PLMA file had a space at the end of them, which made them longer than other lines. This problem was hard to locate initially, but once the lines were identified, fixing them was easy.

As it stands now, AUTOMATCH is somewhat difficult to use. Although it requires many more preparatory steps than RECLSS, AUTOMATCH allows the user to easily control its operation. Matching with RECLSS is currently performed in Headquarters using a predetermined set of matching parameters. It is possible to change these parameter; however, they are generally left at their default values. With AUTOMATCH, the matching parameters can easily be altered as needed for different scenarios. Development of good front and back user interfaces and default matching parameters should greatly simplify the use of AUTOMATCH.

Software Problems

During the testing process a few problems were identified with AUTOMATCH. These problems occurred when reports or extracts of the data were created. The first problem was that the pointer file did not always reference the correct observation. A few times two dissimilar records were reported as duplicates. These records were from different blocks, and would not have even had the opportunity to be linked. The true duplicate was one observation down in the list sequence from the one reported. A second problem was that at times a record was a duplicate on the list, and reported as such on the match files but not written to the report. When the duplicate pairs were written to the report, the master record appeared, but the line for the corresponding duplicate was

blank. We sent MatchWare documentation of these problems. They responded, saying they would correct the problems and send us a new version of AUTOMATCH.

One problem was also noted with the RECLSS system. In a few cases the mainframe would identify two names on the PLMA list that were duplicates but non-matches with the list frame. These names were listed together on the match files, but neither of them appeared on the nonmatch file. Consequently, the potential new operation was overlooked and not added to the list frame.

CONCLUSIONS

The results of this comparison suggest that AUTOMATCH does indeed perform substantially better than the RECLSS system as it currently is used. The number of false matches and false nonmatches in AUTOMATCH are both lower than the RECLSS system. AUTOMATCH made 61 false matches compared to 855 false matches made by RECLSS. AUTOMATCH made 575 false nonmatches compared to 1,533 with RECLSS. Overall, AUTOMATCH's false match rate was 4.6% lower than that of RECLSS and its false nonmatch rate was 12.2% lower.

There are several reasons for the differences in the two systems. The primary reason is that a clerical review was performed with AUTOMATCH and no clerical reviews were done with RECLSS. As decisions were made regarding the clerical record pairs, they were considered the truth. This improved the accuracy of AUTOMATCH. The clerical review process is crucial if optimal results are to be obtained. It was especially impor-

tant for this test because the PLMA file had so few discriminating variables. All clerical review cases in RECLSS were considered matches. This inflated its false match percentage. A comparison of doing a clerical review between passes and after all passes have been run seems to indicate that performing a single clerical review after all passes have been run will work.

Another reason for the differences in the two programs was that the parameters used were not the same. Although the programs are based on the same fundamental statistical theories, the matching algorithms, probabilities and thresholds were not the same for both. Different parameter values led to different component weights being calculated for each comparison. In turn, these were then compared with different threshold values magnifying the differences in the two programs.

Development of default parameter file and matching specifications for AUTOMATCH is an area where further study is needed. The master data file created with the results of this study will be used for this research. After conducting independent research, the Bureau of the Census has found that names and addresses from Ohio can be used to develop matching specifications which are effective throughout the United States [7]. Therefore, most default parameters files developed using the Ohio master PLMA file can be reasonably applied to all other states.

Because the PLMA file only had names and addresses, it is an example of a file that would fit the worst case scenario. Most new source files contain at least phone number, social security number, or another identifying variable besides name and address. Using

the PLMA file as a master file for future developments will be advantageous because its characteristics will be applicable to most new source files.

RECOMMENDATIONS

Despite the fact that the error rates for AUTOMATCH were lower than those of RECLSS, AUTOMATCH has several areas where improvement is needed before it is made fully operational. As it stands now, AUTOMATCH is still complicated to use. It is especially critical that the final AUTOMATCH refinements are implemented with the addition of the Census of Agriculture to the NASS program. The following are some suggestions of ways to make AUTOMATCH an easier and more accurate system.

1. *Good Front and Back User Interfaces.* Good front and back user interfaces need to be developed to make AUTOMATCH more user friendly. This development will greatly reduce the time required to process a match and review its output. Ideally this would be in a Windows format and look similar to other software programs used by NASS. Development of these interfaces is currently underway.

2. *Clear Documentation.* Along with a well-designed software package, clear documentation is crucial to the success of any program. One of the reasons RECLSS is perceived as a black box by many users is that there is currently very little available documentation for the program. Development and maintenance of both a written manual and online help will prevent users from also perceiving AUTOMATCH as a black box.

3. *Record standardization problems.* The standardization portion of AUTOMATCH appears to be functioning well. The Texas and California list frames were both standardized as a test of the system, and relatively few problems were encountered. In order to continuously improve the process, a list should be kept of records that do not standardize properly. This list can be used to learn where future improvements in the standardization code files should be made.

4. *Default matching specifications, thresholds and parameter values for commonly used files.* Processing a match through AUTOMATCH requires many more preparatory steps than RECLSS. The combination of good front and back user interfaces combined with default matching parameters will simplify the matching process with AUTOMATCH. Standard parameter files should be developed for common new list sources and unduplication of the master list frame. This would allow AUTOMATCH to be operated without requiring users to learn a large amount of background statistical theory. It would also standardize procedures throughout the State Offices. Having the master PLMA file will make this task of developing default parameters more effective.

As these standard schemes are developed there are a number of questions that need to be answered. They include things such as:

What is the optimal number of blocking variables?

Which variables are effective for blocking?

Which variables are best to include in the match?

What is the effect on matching accuracy if we eliminate clerical reviews between passes, and only review the clericals after all passes are run?

5. *Clerical Review*. It is critical that a review of clerical records is conducted if good match results are to be obtained. This will require more work of the State office employees, but will lead to more desirable results. The clerical review portion of the software is one area that especially needs revision. As it stands now, only one person can operate the clerical review program at a time. It is not easy to access records, or review previous decisions. In order to keep a record of decisions made about records for this research project, we created our own format and printed the clerical matches. We then looked through stacks of output to resolve the cases. The clerical review process is one of the primary characteristics of the old system that we wanted to improve. Reviewing the output on the screen is much more manageable and efficient than paper.

If AUTOMATCH is going to be made fully operational, it is very important that the communication level between the users, the operational programmers, and the researchers remains high. This will allow the system to continually be improved.

REFERENCES

- [1] Fellegi, Ivan P. and Sunter, Alan B., "A Theory for Record Linkage," *Journal of the American Statistical Association*, Vol. 64,1183-1210, 1969.
- [2] Day, Charles, *Record Linkage I: Evaluation of Commercially Available Record Linkage Software for Use in NASS*, Survey Technology Branch, Research Division, National Agricultural Statistics Service, USDA, 1995.
- [3] Day, Charles, *Record Linkage II: Experience Using AUTOMATCH for Record Linkage in NASS*, Survey Technology Branch, Research Division, National Agricultural Statistics Service, USDA, 1996.
- [4] National Agricultural Statistics Service (1979), "List Frame Manual," U.S. Department of Agriculture, Washington D.C.
- [5] Jaro, Matthew A., AUTOSTAN Generalized Standardization System, Version 3.4, MatchWare Technologies Inc. 1994.
- [6] Jaro, Matthew A., AUTOMATCH Generalized Standardization System, Version 4.0, MatchWare Technologies Inc. 1996.
- [7] Winkler, William E., U.S. Bureau of the Census, Personal Conversation, July 1996.