United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research Division

STB Research Report
Number STB-96-01

March 1996

Record Linkage II: Experience Using AUTOMATCH for Record Linkage in NASS

Charles Day

RECORD LINKAGE II: EXPERIENCE USING AUTOMATCH FOR RECORD LINKAGE IN NASS, by Charles Day, Technology Research Section, Survey Technology Branch, Research Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC 20250-2000, March, 1996, Report No. STB-96-01.

ABSTRACT

NASS uses record linkage for reducing the presence of duplicate names on its list sampling frame of farm operators and agribusinesses. In the late 1970's, NASS developed an automated record linkage system which runs on an IBM mainframe for this purpose. With changes in technology and tightening budgets, the need has arisen for portability between platforms, integration with client/server technology, interactive operation, and reduced resource expenditures. The availability of commercial record linkage solutions has made the development of a new system or an expensive and difficult rewrite of the old system unnecessary. This report summarizes NASS's experience to date with one such commercial package, AUTOMATCH.

The paper begins with a brief history of the project, a description of AUTOMATCH, its companion software, AUTOSTAN, and an explanation of how they are used to do record linkage. Next, the report presents a section that describes three matches, which test AUTOMATCH's fitness for particular record linkage tasks that NASS does. Brief descriptions of the other matches we have undertaken come after that section. Finally, we conclude that AUTOMATCH is suitable for NASS's needs, and that additional work needs to be done to develop front- and back-end software and default parameter files for common operations.

KEY WORDS

Record linkage; List sampling frame; Duplication; Software; AUTOMATCH

ACKNOWLEDGEMENTS

The author would like to thank the members of the List Sampling Frame Section and Guy Pense of the Systems Services Branch for their help in evaluating AUTOMATCH; Dale Turnipseed of the Texas SSO for evaluating the Texas unduplication results; Mike Fleming and Orrin Musser of the Survey Research Branch, Robin Roark of the Commodity Survey Section, Jim Cox of the Washington SSO, and Perry Poe of the California SSO, for allowing us to use AUTOMATCH in support of their projects; and Dave McDonell and Robyn Thoelke of the Systems Services Branch for help in developing supporting software. Thanks to management (Ron Bosecker, George Hanuschak, and Roberta Pense) for their support of this project.

Table of Contents

SUMMARY

The National Agricultural Statistics Service (NASS) uses record linkage to minimize duplication of farm operators or operations and agribusinesses during construction and maintenance of its list sampling frame (list frame). In 1992, NASS decided to replace its current list frame database, the Real Time Mail Maintenance System (RTMMS) with a new database, the Enhanced List Maintenance Operations (ELMO) system. This decision caused a reevaluation of the current record linkage system. The need for portability between platforms, integration with client/server technology, and interactive linkage, combined with a desire for a system which required fewer personnel resources to operate, led to the decision to explore new record linkage solutions. The current record linkage system, developed in the late 1970's, required over 50 staff years to produce. Fortunately, in the intervening 15 years, several software packages that do list unduplication have become available commercially, eliminating the need to repeat such a costly development effort or to attempt to port the complex set of legacy COBOL and FORTRAN programs that comprise the current system to a new platform. The purpose of this project is to investigate the availability of commercial packages and to propose one for adoption as the core of a new record linkage system. This report summarizes NASS's experience to date with AUTOMATCH, the package chosen for evaluation.

AUTOMATCH, a product of MatchWare Technologies, is a generalized record linkage solution, meant to do automated record linkage for many different applications. The user specifies a match by coding parameter files. The software is complete, in the sense that the user can do all parts of the matching process without any additional programs. (However, NASS needs to develop additional software to enhance functionality and ease of use within NASS.) AUTOMATCH comes with companion software, AUTOSTAN, which does standardization, the process by which free-formatted name and address fields are broken into their component parts, so that these parts might be compared meaningfully between records. AUTOSTAN is very powerful and highly customizable.

This report describes three matches in detail, each done to test AUTOMATCH's suitability for a specific task. The first of these is a match between the June Agricultural Survey area frame sample for Wyoming for 1993 and the Wyoming list frame. This match was done to test AUTOMATCH's ability to detect overlap between the area frame sample and the list frame. AUTOMATCH successfully detected 94 percent of the overlap, but missed two extreme operators. (Human review detected 99 percent of the overlap, including all of the extreme operators.) This result showed that, while AUTOMATCH would be useful in rapidly detecting most of the overlap, it needs to be supplemented by human review of extreme operators.

The second match was a linkage between the North Carolina list frame and a new list source. The match was done once using the current, mainframe system and then again using AUTOMATCH to compare the results between the two systems. There are two types of errors in record linkage. One is a false link, the association of two records that do not represent the same unit in the population; the second is a false nonlink, the failure to associate two records that do represent the

same unit in the population. AUTOMATCH outperformed the current system as it is now used. AUTOMATCH had a false match error rate of 1.1 percent; the current system had a rate of 4.6 percent. AUTOMATCH had a false nonmatch error rate of 4.9 percent; the current system had a rate of 8.7 percent.

The third match described in detail was an unduplication of the Texas list frame. The Texas list was chosen because it is NASS's largest list frame. AUTOMATCH had no problem handling a file of this size. The results were evaluated by Texas's list frame statistician. After his evaluation, AUTOMATCH found an overall duplication rate of 0.36 percent. This result was consistent with independent research estimates of list frame duplication in Ohio and North Carolina [1].

This report then discusses five other matches in less detail. Besides allowing us to test new matching and standardization approaches, these matches also provided us with some insight into advantages and difficulties of using AUTOMATCH. In particular, we concluded that NASS could use AUTOMATCH in several applications for which the current system is poorly suited. One such application would be to match administrative data files to sample files to enable the use of administrative data in estimation. Also, we discovered that people who do not have the time to become record linkage and AUTOMATCH experts find the software somewhat difficult to use as it comes from MatchWare.

This finding led us to make several recommendations about developing front end software and a better possible match review program. We also recommended that NASS develop a capability to store possible match resolution information across linkages, along with supplementary programs to do several NASS-specific tasks. Many of these recommendations have already been adopted by the ELMO2 Steering Committee. Our earlier paper, *Record Linkage I: Evaluation of Commercially Available Record Linkage Software for Use in NASS*, discussed our reasons for selecting AUTOMATCH over other commercially available record linkage software. In this paper, we briefly discussed the alternatives of porting the current system to a new platform with added functionality or writing a new, proprietary system. We concluded that both alternatives offer no technical advantage over the adoption of AUTOMATCH as the core component of a new system, and would also be more costly.

AUTOMATCH performed well as record linkage software. We concluded that AUTOMATCH should be the agency's choice as a core component for a new NASS record linkage system.

This report is not a general evaluation of AUTOMATCH. We based our evaluation of AUTOMATCH on NASS's requirements alone. The recommendation of AUTOMATCH is not an endorsement of AUTOMATCH for any other uses outside NASS.

INTRODUCTION

NASS uses record linkage to construct and maintain its list sampling frame (list frame) of agricultural operators and operations with a minimum of duplication. In 1992, NASS decided to replace its current list frame database, the Real Time Mail Maintenance System (RTMMS) with a new database, the Enhanced List Maintenance Operations (ELMO) system. This decision resulted in discussion of the need to replace the current record linkage system. The need for portability between platforms, integration with client/server technology, and interactive linkage, combined with a desire for a system which required fewer personnel resources to operate, led to the decision to explore new record linkage solutions.

The current record linkage system, created in the late 1970's, required over 50 staff years to develop. Fortunately, in the intervening 15 years, several software packages that link records between files and find duplicate records within a file (that is, "unduplicate" the file) have become available commercially. Their availability eliminates the need to repeat such a costly development effort, or to attempt to port the complex set of legacy FORTRAN and COBOL programs that comprise the current system to a new platform. The purpose of this project is to investigate the availability of commercial packages and to propose one for adoption as the core of a new record linkage system. This report summarizes NASS's experience to date with AUTOMATCH, the package chosen for evaluation, and its companion software for standardizing names and addresses, AUTOSTAN.

The paper begins with a brief history of the project and the different linkages we did to test the software. It follows with a description of AUTOMATCH and AUTOSTAN, and how a user does a linkage with them. Next, the report presents a section that describes three of the linkages, each chosen to test AUTOMATCH's fitness for a particular type of record linkage task that NASS does. After that section comes a series of brief descriptions of the other matches that we have undertaken using AUTOMATCH. Finally, the paper presents conclusions and recommendations.

A Brief History of the Project

This project began in the summer of 1992. The priority was to understand how the current system worked and what functionality would be required of the new system. Following a review of documentation for the current system, Technology Research Section staff drafted a set of requirements. The List Frame Section, six State Office list frame statisticians, the ELMO team, and several members of the team that developed the current system reviewed these requirements. Concurrently with the development of the requirements document, we made an active search for record linkage software packages.

The packages fell into three broad categories. The first category is software routines that do specific parts of the record linkage task. Building a system from them requires many additional resources. The second category is specialized mailing list unduplication software that is an integrated part of a direct mail management system. These packages use nonprobabilistic, ad hoc linkage methodologies better suited to their intended purpose than to the construction of a NASS system. The third category is generalized systems.

These systems provide a complete solution for record linkage and use the statistically defensible Fellegi-Sunter record linkage methodology. A user can do linkages with them as they come from the developer. They require the fewest additional resources to build a record linkage system that meets NASS's requirements. AUTOMATCH falls into this last category.

The first report issued from this project, *Record Linkage I: Evaluation of Commercially Available Record Linkage Software for Use in NASS* [2], contains detailed reviews of six commercially available packages. That report also contains background material on record linkage and NASS. **Readers not familiar with NASS and record linkage should read that report first.** The review of the six packages made it clear that AUTOMATCH, from MatchWare Technologies, merited hands-on evaluation. It did not appear worthwhile for NASS to commit the resources to do a hands-on evaluation for any other package.

Testing Approach

Once AUTOMATCH was chosen, we had to decide how to test it. There were two approaches. One was to develop small test files of records for which the true matching results would be known. Developing even small files of this type by hand would have been very resource intensive. The second approach was to do several test linkages, at least one of which would be compared to a linkage of the same two files by the current, mainframe record linkage system. This approach had several advantages. It gave us the chance to do several different types of linkages (for example, within-list unduplications, overlap checks, matching new sources). It also provided an opportunity to learn how well AUTOMATCH met different challenges (for example, large files, sparse information, "messy" name and address fields). For these reasons, we chose the second approach.

Some linkages were done purely for record linkage research purposes. Others were done to aid other research projects or to aid the operational programs of the agency, but all of them taught us something valuable about using AUTOMATCH. When list frame files were needed, we extracted them from the name and address master files on the RTMMS. We used the RTMMS because ELMO was not yet available.

For our first test linkage, we chose to assist a research project to improve NASS's understanding of the nonoverlap domain and problems in sampling from that domain. We used AUTOMATCH to link area frame and list frame records. This linkage tested AUTOMATCH's ability to do overlap checking.

Next, we did a comparison between AUTOMATCH's results and those of the current system in linking a new source file to the North Carolina list frame. This linkage tested the AUTOMATCH's ability to reduce duplication when adding a new list source to the list frame.

When we presented the results of our North Carolina investigations to the Information Resources Management Review Board, they suggested that we test AUTOMATCH's within-file unduplication capabilities. We chose Texas for this linkage because they had the largest state list frame (about 100,000 active records). This linkage verified AUTOMATCH's ability to handle NASS's

largest state list frame and to detect duplication on NASS list frame files.

Besides these three linkages, we have done five other linkages which are described in less detail. Beyond the work documented in this paper, Kara Broadbent of the Applications Research Section (in the Ohio SSO) has done two linkages with the software, and has done important work in optimizing the name standardization process. Her work will be presented in a forthcoming report [3].

BACKGROUND

General Description of the Software

The software is available for mainframes, UNIX minicomputers and workstations, and IBM-compatible microcomputers (with the DOS, OS/2, or Windows NT operating systems). We used the DOS version for testing. (It was least expensive, and MatchWare assured us that their software has the same features across all platforms on which it runs.) It has run well on a Compaq Deskpro 50M (486 DX2-50, EISA bus architecture) with 8 Mb of memory and a 340 Mb hard drive. An additional 1 Gb SCSI hard drive has been used for linkages involving larger files (the Texas unduplication work in particular). The software itself requires less than 5 Mb of hard disk space. MatchWare says that linkages of files of as many as two million records are regularly done on UNIX systems, and that larger linkages would be possible if the system resources were available. The use of pointer files and indexing instead of sorting data files eases the linkage of larger files. It is possible to specify a subset of records in a data file for linkage.

Testing has all been done in batch mode; however, interactive versions of AUTOSTAN's standardizer and the AUTOMATCH modules are available as Windows DLL's and UNIX callable libraries. Sybase Corporation, the manufacturers of the database software that ELMO uses, itself uses AUTOMATCH callable libraries to detect duplication in its customer database.

The completeness of AUTOMATCH means that the cost of developing a specialized system to meet NASS's needs will be less than any of the other options. It also means that improvements to the system can be incremental, as resources become available.

AUTOMATCH and AUTOSTAN came with complete, easily used documentation. Matt Jaro, developer of the software, gave about eight hours of on-site introductory training as part of the software purchase. With this introduction, the user's manual, and free telephone support (which we used liberally), we have accomplished everything we wanted to do with the software.

The software is still evolving. Over the course of the two and one-half years we've had the software, MatchWare has made many improvements, some at our urging. MatchWare has improved (by an order of magnitude) the speed of the indexing program that is run before a linkage. (This is done instead of sorting files.) They have made a similar improvement in the speed of the program that counts frequencies for the computation of u-probabilities. They have also added several additional comparison types, the ability to set weights manually, optional treatment of missing variables, the ability to specify "must agree" variables, routines for generating NYSIIS (phonetic) codes, additional match types, array-type variables, and improved clerical review software. The next

version promises to include improvements that will speed input/output by reducing the amount of time spent "seeking" records on the hard disk.

As of February 1996, other government users of AUTOMATCH include the Bureau of Labor Statistics and the Internal Revenue Service. Nongovernmental users include the United Parcel Service (UPS).

An Overview of the Record Linkage Process

Diagram 1 gives an overview of the record linkage process using AUTOMATCH and AUTOSTAN. Diagrams 2 through 5 further explain each step in Diagram 1. These diagrams assume that a list frame file is being linked to a new source file of names. The process is similar for an unduplication, with the modification that only one file needs to be standardized, one dictionary prepared, and the list frame file is linked against itself rather than against another file. Before beginning the linking process, the user must get the file or files to be linked into ASCII format, with the location and contents of each field defined.

A Description of AUTOSTAN and Its Use

The first step in the linking process is the verification and standardization of names and addresses. Standardization is a key task in doing record linkage using names and addresses. It is the process of identifying all of the parts of a name or address, parsing the name or address into its various parts, and substituting standard words for nonstandard abbreviations and nicknames. This is done so that like parts of the name and address in two records may be compared to each other, despite the original format of the fields in the two records. AUTOMATCH has companion software, called AUTOSTAN, which does this task. Diagram 2 shows the steps we used in verifying place names and ZIP Codes, and standardizing the names and addresses in the files to be linked. AUTOSTAN can handle variable record length, multiple line, and delimited input files, besides fixed field, fixed record length files. The software can produce a fixed field, fixed record length file from any of these formats. This is important, since AUTOMATCH requires fixed record lengths and fixed fields.

The first step we undertook was the verification of the accuracy of the place names, states, and ZIP Codes in the addresses. This is the one step which AUTOSTAN could not do. We wrote a program, in C, called Plverify to do this, using a "dictionary" of place names, with their standard abbreviations, associated range of valid ZIP Codes, and latitudes and longitudes. This dictionary is a reformatted version of the same one used in the current record linkage system. For each record in the input file, the program verifies that the place name exists, that the ZIP Code is in the valid range. The program then writes out the original record, preceded by the standard abbreviation for the place name, a valid ZIP Code, and the latitude and longitude. If the place name is not in the dictionary, the program writes out the unverified place name, state, and ZIP Code with blanks in the latitude and longitude variables. An error message is written to a file if the place name is not in the dictionary, or the ZIP Code or state is incorrect.

Next, we used AUTOSTAN to standardize the names and addresses on the two files. AUTOSTAN uses a "process" for each type of field it standardizes. Each of these processes consists of three parameter files. The

Diagram 1.--Overview of Record Linkage Using AUTOMATCH and AUTOSTAN

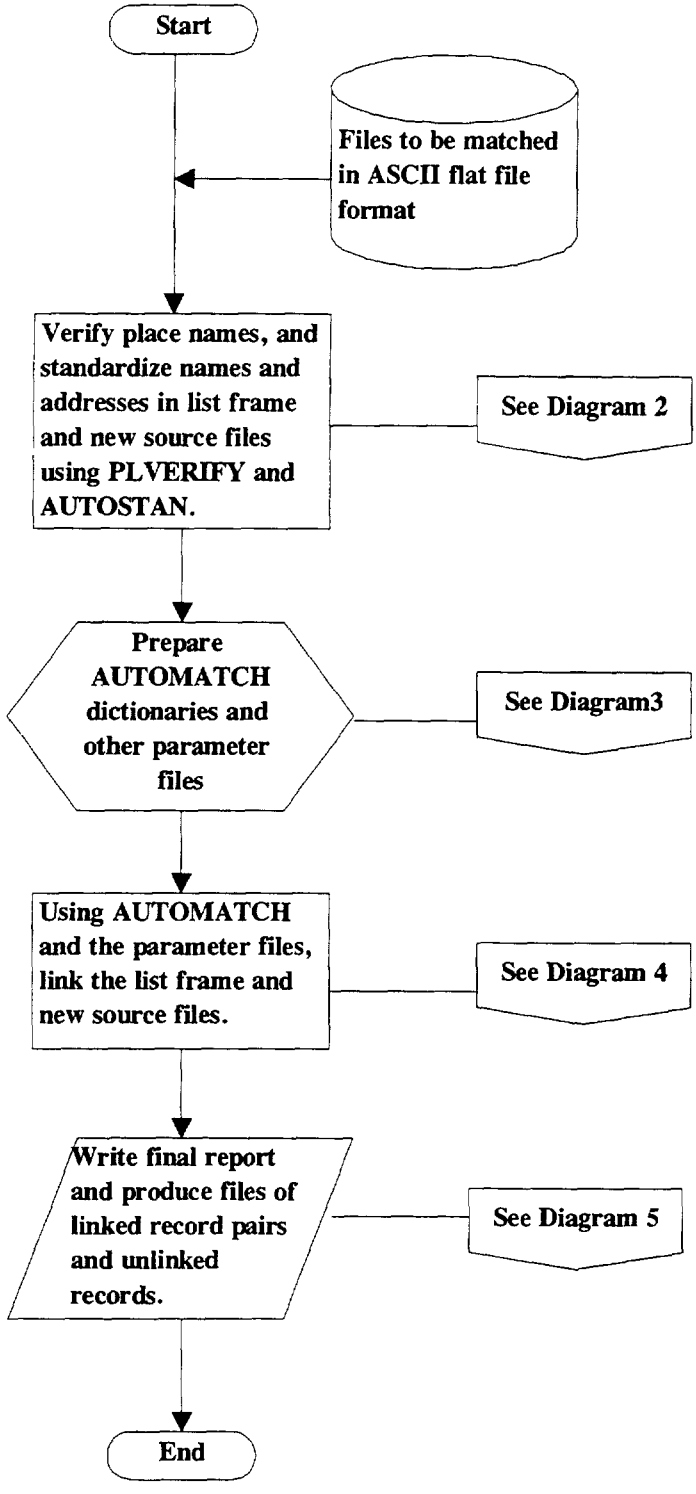
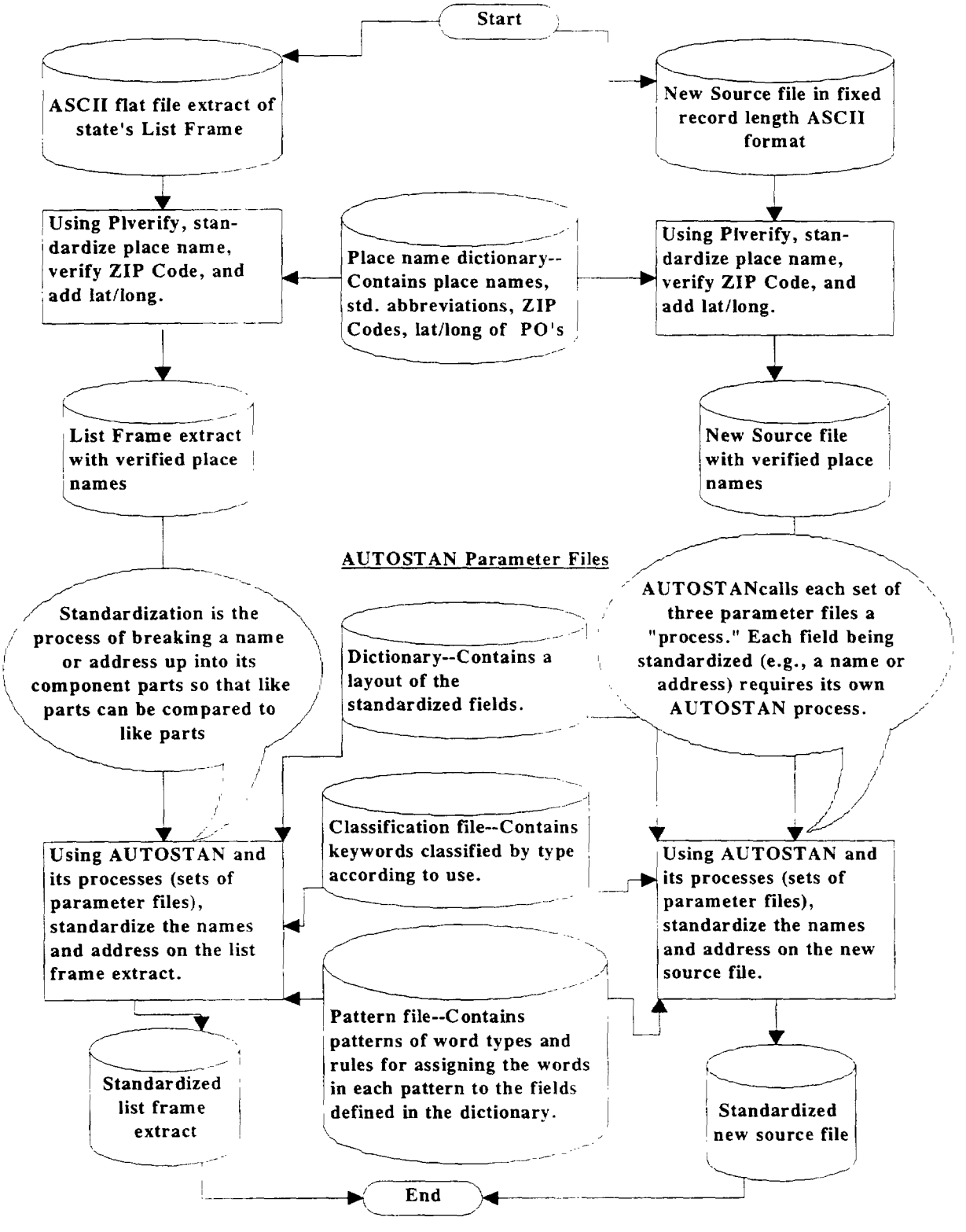


Diagram 2.--The Standardization Process



first, the "classification" file, contains keywords and a type for each keyword. For example, the classification file might classify "Mary" as a given name, "Slaughter" as a surname or nonperson name, and "Plant" as a surname or nonperson name.

The second file, the "dictionary" file, specifies the parts into which the field is to be parsed. For example, a name field might be broken down into a title, a first name, a middle name, a surname, and an operation name. The output record from AUTOSTAN is a fixed length, fixed field record containing these standardized fields, followed by the original input record.

The third file, the "pattern" file, contains rules for parsing different patterns of word types. For example, one rule might say that if an input line consists of a given name followed by a surname or nonperson name, then the given name is placed in the first-name field, and the surname is placed in the surname field. A second rule might say that if the input line consists of any number of unidentified words followed by two surname or nonperson name words, then the input line is placed in the operation name field. In the example above, "Mary Slaughter," the name "Mary" would be placed in the first name field, and "Slaughter" would be placed in the surname field. On the other hand, the input line "Hanuschak Slaughter House" would be placed in the operation name field. In these two examples AUTOSTAN would have used context to treat the word "Slaughter" properly.

When we began a linkage, we used a text file viewer (in our case, "list.com") to examine name and address fields for any name or address forms which we knew AUTOSTAN would not standardize properly. A less experienced user could make a preliminary AUTOSTAN run and examine the output with a text file viewer for names or addresses which did not standardize properly. If an unacceptable proportion of such forms were present (say, more than 1 percent of the fields), we made modifications to the name and address standardization processes to handle them. This was done by adding words to the "classification" files, or adding new rules to the "pattern" files.

We have primarily used two processes. We call them our default processes. We created the process to handle name fields. Using a dictionary of keywords developed for use in the current record linkage system, we created a classification file for names. After examining the different parts into which the current system parses names, we wrote a dictionary file. Finally, we developed a pattern file to specify the rules for standardizing names.

The name process breaks up a free formatted name field into as many as 16 parts (although any one name will not have all 16 parts). This is necessary so that the matching software can compare like parts of a name to like parts (for example, surname to surname, title to title, operation keywords (for example, farm, incorporated) to operation keywords). Many records that NASS attempts to link have few identifying variables (like name, address, phone number). Dividing the variables into as many meaningful component parts as possible uses the information that is present in the record as efficiently as possible.

The second process, which MatchWare created, handles addresses. The address process operates similarly to our name

A Short Glossary of Record Linkage Terminology[1]

Blocking--When linking two files, A and B, it would be ideal to examine all of the pairs of records containing one record from file A and one record from file B. Unfortunately, this is usually impossible, due to the number of comparisons which would have to be made. Instead, record pairs are compared only within *blocks* in which all records have the same value for some variable or variables, referred to as *blocking variables.*

m- and u-probabilities--The Fellegi-Sunter record linkage theory used by AUTOMATCH is based on probabilities. The theoretical definition of these probabilities is covered in *Record Linkage I: Evaluation of Commercially Available Record Linkage Software for Use in NASS.* The m-probability for a particular value of a variable depends on its accuracy. The u-probability depends on its discriminating power, which is a function of the frequency of occurrence of the value.

Match weight--In the Fellegi-Sunter theory, decisions about whether or not a pair of records represent the same unit in the population are based on comparisons of match weights to a *weight cutoff.* The match weight for a pair is the sum over all of the linking variables of the $\log_2$ of the ratio of the m- and u-probability for the outcome of the comparison of the variable's value(s) on the two records. In simpler terms, match weights reflect, in a rigorous way, the likelihood that two records being compared represent the same unit in the population (that is, that they match). Higher weights reflect a greater likelihood of matching.

Record linkage--The association of records representing the same unit from one or more files representing the same population by comparing identifiers or *linking variables* (like name, address, Social Security Number) for construction and maintenance of a master file for a population.

[1]A more extensive glossary, along with an explanation of the Fellegi-Sunter record linkage theory, are offered in *Record Linkage I: Evaluation of Commercially Available Record Linkage Software for Use in NASS.*

process. It identifies as many as 14 parts of an address.

Over the course of our research, we have identified many difficult name and address forms. When we felt that these were likely to recur, we incorporated modifications into the default processes to handle them. MatchWare also continues to do this for the address process. In this way, the default processes have evolved to handle more name and address forms. We expect that the default processes will perform adequately when the software becomes operational, but the capability to make occasional modifications ensures that the software can handle new demands in the future.

While we believe that default processes can handle standardization for linkage purposes, it is possible that specialized processes could be created to handle other functions. One

example is postal standardization. While the current address standardization process is specialized for record linkage, there is no reason a process could not be written to create a postal standard address.

Besides the files defining the processes, one additional parameter file is needed to run AUTOSTAN, the command file. This file contains the names of the input file, the output file, the input record size, the input file type, and the names of the processes to be used with the starting position and length of the field to be standardized for each process.

Once the parameter files were modified, AUTOSTAN was invoked and the standardized files were produced.

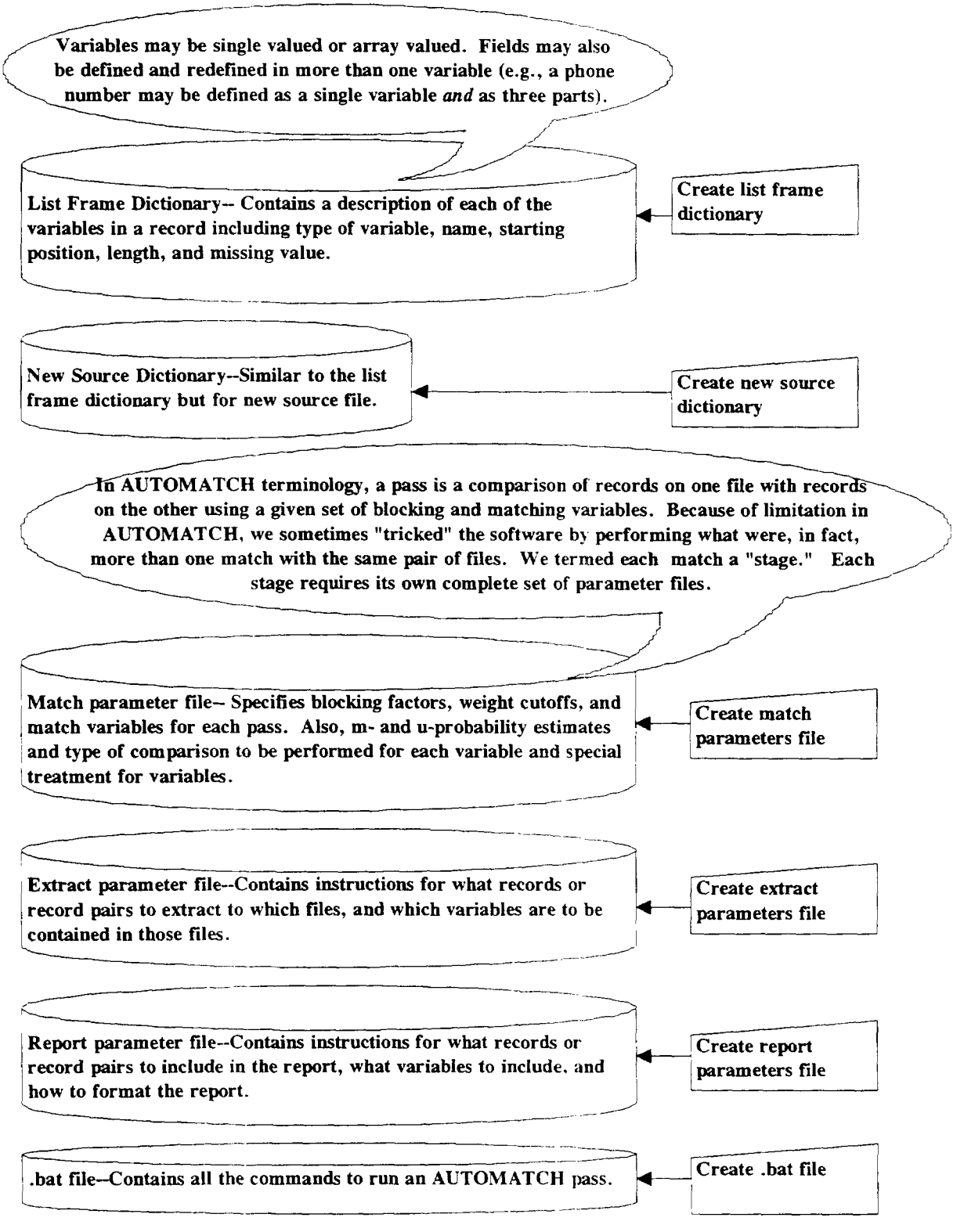
A Description of AUTOMATCH and Its Use

AUTOMATCH is a generalized record linkage system, meant to do linkages for many purposes without additional programming. The software uses the statistically defensible Fellegi-Sunter record linkage theory, as does NASS's current system. It can unduplicate a single file, link two data files, and link a data file to a reference file (like a file of geographic codes) to attach additional items to the data file.

We adopted three terms, "match," "pass," and "stage," to describe our linking schemes. The entire operation of linking two files is called a match. Each time the two data files are compared to each other using a particular blocking scheme and set of linking variables is called a pass (because it represents one "pass through the data"). These are AUTOMATCH terms. We invented the term "stage" because AUTOMATCH does not allow passes without linking variables (these passes are effectively an unblocked linkage on the nominal blocking variables) to be mixed with passes which have linking variables. This limitation forced us to extract the unlinked records after each stage so that we could use them as input to the next stage. We never used more than two stages in a match. Figure 1 on page 18 shows one example of stages and passes for a match.

Preparing to use AUTOMATCH.--Diagram 3 shows the parameter files used by AUTOMATCH to give the user control over the linkage process. The user specifies the structure of each data file with a dictionary file. In AUTOMATCH terminology, a dictionary is a file that contains the name of a data file, its record length, and a description of each variable on its records. A variable is defined by giving its name, starting position within the record, length, and missing value. (The missing value is defined as zeroes or spaces. Whenever this value is detected, the variable is considered to have a missing value, as opposed to a true zero or blank as a value.) The AUTOMATCH dictionary file reflects the format of the input data file after standardization by AUTOSTAN (that is, with the standardized name and address components). A physical field on the data record can be contained in more than one variable. For example, if the first ten bytes of a record contain a telephone number, the dictionary may contain one variable, phone number, which is defined as the first ten bytes of the record. Either alternatively, or in addition, three other variables can be defined from the same field. These might be called area code, which would be defined as the first three bytes of the record, exchange, which would be defined as bytes four through six, and ran_digits, which would be defined as bytes

Diagram 3.--The Parameter File Creation Process for AUTOMATCH



seven through ten. We often do this so that we can treat several linking variables as a single variable when defining a report (review) or extract format.

After preparing a dictionary for each file, we prepared a match parameters file. The match parameters file specifies the type of match to be done (unduplicate a single file, link two files, or link to a reference file); the blocking and linking variables for each pass; treatment of missing values; weight cutoffs for each pass; the estimated m- and u-probabilities; optional variable types (such as, "no frequency count," or "critical"), and types of comparison for each linking variable. AUTOMATCH allows several different types of comparisons, including character for character, information-theoretic text string comparison, and a special comparison for geographic coordinates.

AUTOMATCH also allows the user to define array-valued variables; that is, a single variable name with multiple values. These values may be located adjacent to each other or physically separated within the record. When an array variable is used in a comparison, the software compares all of the values of the variable on one file to all of the values for that variable on the other file. This is especially useful for situations where one or both files contain multiple phone numbers or names.

To specify the match parameters file, we first examined all of the identifiers present in both files (variables like name, address, SSN, EIN, telephone number), and developed a strategy for using them in the linkage. Such a strategy outlines which variables should be compared, how many passes to make through the data, which variables to use for blocking, what type of comparison to use in each pass for each variable, and how variables should be defined (for example: a phone number can be broken down into area code, exchange, and the last four digits; or, if there are two phone numbers, they can be treated as a single array-valued variable for linking purposes).

The creation of the parameter files assumes knowledge of both the rudiments of the Fellegi-Sunter theory and AUTOMATCH. The parameter file preparation process is lengthy, and the objection has been raised that a state list frame statistician would not have time to go through it. Unfortunately, there is no "front end" provided with AUTOMATCH which automates the creation of parameter files. We recommend that NASS develop such a front end. Also, we believe that for the most common list frame applications, the agency can prepare standard parameter files that will require few, if any, changes from linkage to linkage. We do recommend, however, that the performance of these defaults be reviewed regularly, and that they be adjusted periodically to maintain quality. Weight cutoffs, m-, and u-probabilities can be estimated using utilities included with the package.

Besides the dictionaries and match parameters, parameter files are coded to specify the creation of extract data files after the linkage and to specify any reports that are required from the linkage.

After creation, the match parameters and dictionary files are compiled, and the compiled versions are used to gain efficiency. If any of the files are subsequently modified (particularly if weight cutoffs are changed in

the match parameters file), they must be recompiled.

Doing a linkage with AUTOMATCH.-- Diagram 4 describes the process of using AUTOMATCH to link two files. Once the files to be linked have been through the standardization step, and the parameter files have been coded, the next step is to do the first pass of the match.

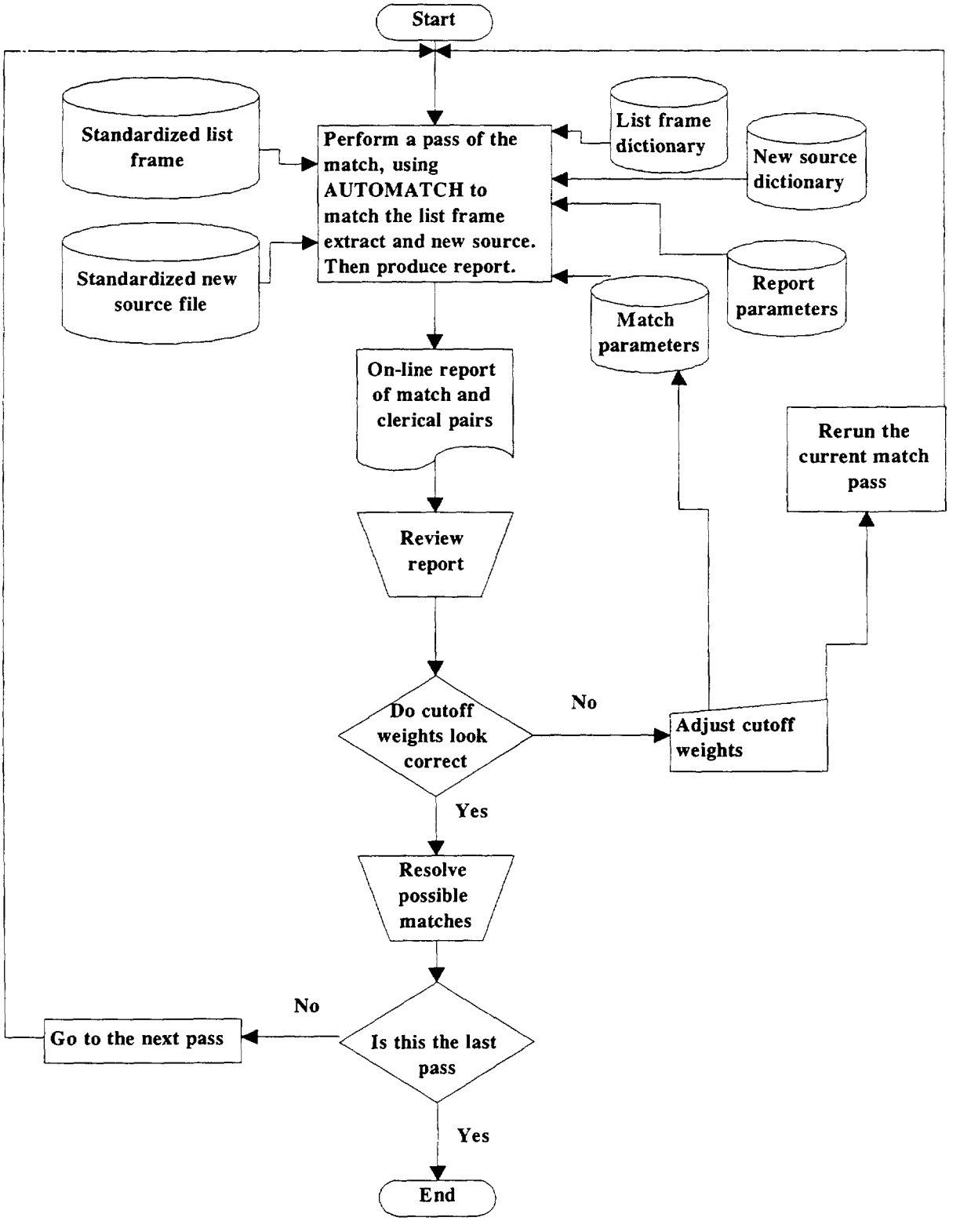
At this point, some parameters, like weight cutoffs and m-probabilities are estimates, based on experience. If the files being linked are very similar to others linked in the past, these estimates may be adequate without adjustment. Like all Fellegi-Sunter linkage systems, AUTOMATCH produces a set of nonlinked records for each input file, a set of linked record "pairs" (one record from each file, plus any duplicates detected on either file), and a set of possibly linked record pairs. The possibly linked pairs are called "clerical" in AUTOMATCH terminology, and the process of allocating them to either the linked or unlinked set is called "clerical review." After the first pass, a report should be generated, using AUTOMATCH's report generation utility, which shows the possibly linked record pairs.

The report generation utility prepares customized reports, which are formatted as print files, with embedded printer control characters, but are, in practice, reviewed using a text file viewer like "list.com." This utility allows the user to create a report, complete with page breaks, headers, and summary information about each pass, containing any of the variables on the records in the two files being linked. There is one limitation. The on-line clerical (possible link) review software uses the same format used for the report. In practice this has meant that we constructed our report format so that it produced the best format of three 80-character lines for the clerical review program. We have found that, while this is not ideal, it works well enough.

Besides the report generated by this utility, the matching program itself generates statistics on the linking process. By setting a command line switch, different levels of detail can be included in the report. These include such useful information as the m- and u-probabilities for each value of each linking variable; agreement, disagreement, and missing weights for each value; and a histogram of the match weights for all record pairs. The histogram can be useful in estimating weight cutoffs for the linked, possibly linked, and unlinked subsets.

We examined the report and the match statistics, to see that the lower weight linked pairs did not include pairs that represent different units in the population. If this were the case, then the "match" cutoff weight was adjusted up, to exclude the incorrectly linked pairs. Conversely, if the upper part of the "possible match" region contained a range of weights for which all pairs were matches, then the cutoff was adjusted down, to include the pairs in this region. At the other end of the possible match region, we were concerned if we found that all pairs below a certain weight were nonmatches. If this were the case, we adjusted the "clerical" weight cutoff up, to exclude these pairs. If, on the other hand, we discovered that we were still finding many clear matches all the way down to the "clerical" cutoff, we lowered the cutoff, on the assumption that additional matches might have been lost below it.

Diagram 4.--The Record Linkage Process Using AUTOMATCH



When weights were adjusted, we ran the pass again. After some experience, we found that the number of iterations required to find correct cutoffs was reduced, usually to one, as our initial estimates became better. After we ran a pass, and our review of the report of linked and clerical pairs satisfied us that the weight cutoffs were acceptable, we went on to review the clerical cases.

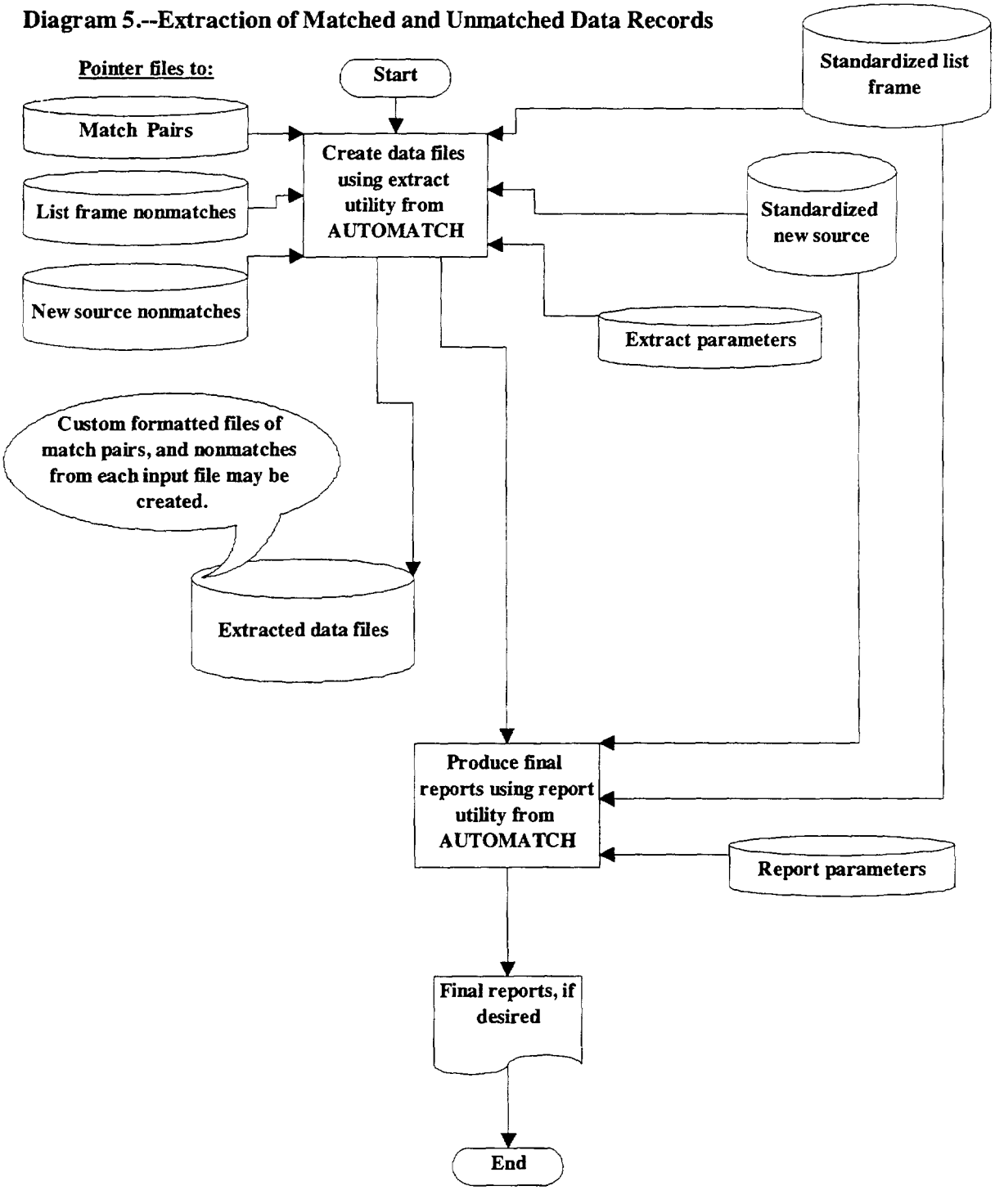
The clerical pairs are available for review after each pass. A review program, whose functionality is sufficient for many linking applications, but marginal for NASS's operational needs, is provided with the software. This program, using a character-based interface, presents a custom-formatted subset of the variables on each record pair (up to 240 characters per record in three 80-character lines). The review program also allows the user to take action on the pair. (For example, the user may decide that a possible link pair should be linked, unlinked, or left as a possible link for review in the future.)

Because reviews are done between each pass, the user may sometimes see the same possible-link record pair on subsequent passes, although he or she had made a decision about the pair following an earlier pass. Reviewing these pairs again would be a waste of time, so the utility allows the user to apply the decision made after the first pass automatically (without having to see the pair again). Previous passes' decisions are applied only to record pairs available for clerical review based on the match and clerical weight cutoffs in the current pass, and only for those record pairs that were clerically reviewed in previous passes. This is a recent improvement in the software, and it reduces considerably the time needed to review possible links. Unfortunately, no such capability is available across linkages. Also, only one reviewer at a time can review possible matches. For AUTOMATCH to become NASS's record linkage software, more functionality needs to be added (see the Recommendations section of this paper); however, this utility was adequate for research purposes.

After we reviewed the first pass's clerical cases, we ran the next pass. The process described above was continued until the last pass in the match had been completed. The current NASS system makes one pass through the data files. AUTOMATCH allows a maximum of eight passes through the data in each stage. Each pass represents a separate attempt to link the two files, using different blocking and linking variables. The software allows multiple blocking variables in a single pass. Only rudimentary options are available to address violations (always present, especially in address variables) of the independence assumption of the Fellegi-Sunter theory. No routine is available for estimating the level of linkage errors.

Extracting Data Files.--Diagram 5 describes the process of extracting data files and producing final reports (listings of matched pairs and summary statistics on the matching process). AUTOMATCH includes a utility for extracting files containing different types of record pairs and unlinked records. The user decides which types of records to include in each file, and what subset of variables from the input data records to include in each extracted record. For example, the user might want a file of all unlinked new source records, and a file of all matched records. The extract utility has proven very

Diagram 5.--Extraction of Matched and Unmatched Data Records



valuable, and is usable by NASS with little if any modification. We have used it to create outputs that could not be created using the current system.

DESCRIPTION OF MAJOR MATCHES

This section of the paper presents detailed results of three matches that test AUTOMATCH. The first is a linkage between the June Agricultural Survey area frame sample for Wyoming for 1993 and the Wyoming list frame to evaluate overlap between the two frames. This match evaluated AUTOMATCH's capabilities as an overlap detection tool.

Next, we linked a new list source and the North Carolina list frame. This linkage duplicated one done with the current mainframe record linkage system. The results of the AUTOMATCH linkage were compared to those from the current system to evaluate the quality of the AUTOMATCH software, and to test AUTOMATCH's ability to update the list frame with a new source.

The third linkage tested AUTOMATCH's ability to do within-list unduplication. We chose the Texas list frame file for this because it is NASS's largest. The Texas list frame statistician evaluated the results of this linkage.

Nonoverlap Domain Study

Purpose--The purpose of this research effort was two-fold. First, the Survey Research Branch (SRB) was engaged in a project to improve NASS's understanding of the nonoverlap domain and problems in sampling from that domain. In order to identify nonoverlap area tracts that represented operations on the list frame that were not classified for the June Agricultural Survey (JAS), SRB needed to link the area sample and the entire list frame (including inactive records). Second, this linkage also evaluated AUTOMATCH as an overlap-detection tool, and provided our first large-scale experience with the software.

Linking Scheme--We linked the area frame records (representing both agricultural and nonagricultural tracts) from the RTMMS Name and Address Master file to the list frame records (both active and inactive) from the RTMMS Name and Address Master file. Each record on both files contained the following common identifiers:

1) nine-digit record id,
2) cross-reference id
   For the area file this field contained the list record to which the area record was overlapped during the June survey (if any). For the list file it contained the list-cross-reference (LCR) (if any),
3) SSN,
4) EIN,
5) telephone number,
6) primary name,
7) secondary name,
8) street address,
9) place name,
10) ZIP Code.

Name and address standardization was carried out as follows:

1) Place names and ZIP Codes were verified using a C-language program developed in-house and a modified version of the NASS place name dictionary for Wyoming (originally developed for use with NASS's mainframe data manipulation program).

2) Addresses were standardized using the AUTOSTAN software package, along with a process developed by MatchWare for the Census Bureau's Geographic Division.

3) Primary and secondary names were standardized using AUTOSTAN along with patterns developed in-house which used a modified version of the NASS name dictionary. All records contained standardized fields for corporation name, partnership name, and the components of individual names. Consider a record with an individual name in the primary name field, and nothing in the secondary name field. The fields for the components of the primary individual name contained the name information, and the corporation and partnership primary name fields were blank. All secondary name fields were blank. This created some problems because AUTOMATCH interpreted the blank fields as "missing," and the early versions of AUTOMATCH assigned a positive weight for the comparison outcome "missing." MatchWare has since changed the software to default to a zero weight in these cases. The user retains the option of adopting the earlier treatment where that would be appropriate.

We tried to link the records using all of the linking variables in a single pass. Due to the sparseness of many highly-weighted variables and the problem cited above with the missing values, this approach was not successful. The sum of the many small, positive "missing" weights (including weights for the "missing" name fields that were an artifact of standardization) was swamping the discriminating power of the variables that were present.

We subsequently adopted a two-stage strategy that called for linking in several passes. The first stage used what we thought were highly reliable variables with considerable discriminating power, such as SSN, EIN, and telephone number. (See the Results subsection for a discussion of telephone number as a linking variable.) A second stage using name and address as linking variables followed. We needed to break the match up in this way because early versions of AUTOMATCH did not allow exact matches on a single variable to be run efficiently in the same match with linkages on multiple variables using other comparison methods. Figure 1 shows the scheme we adopted.

Note that we used two separate stages because AUTOMATCH generated an error message when we combined passes containing linking variables in the same match run with passes that did not use any linking variables. The idea of a pass with no linking variables also needs some clarification. This is essentially an unblocked linkage on the nominal blocking variables. Use of this technique prevented missed EIN, SSN, or Phone Number matches due to records being blocked apart.

The multistage/multipass approach did have drawbacks; the main one being that records linked in earlier stages or passes were not available for linking later. While this prevented the same links from being made in multiple stages or passes, it also prevented a list record that linked to an area tract from being available to link to another tract in a later stage or pass. MatchWare has solved

Figure 1.--Matching Scheme for Wyoming Overlap Study

Stage/Pass	Blocking Variables	Linking Variables
1/1	SSN	None
1/2	Phone number	None
1/3	EIN	None
2/1	ZIP Code	Partnership Name
2/2	ZIP Code	Corporation Name
2/3	ZIP Code	Primary and Secondary Name, Address, Place Name, and State
2/4	Surname Soundex	Primary and Secondary Name, Address, ZIP Code

this problem by adding a new type of match that uses the list frame file as a "reference" file. This makes all of the list frame records available for linking in all stages and passes, no matter whether they have linked earlier.

Results of Matching--Matching divided the area and list files into three groups, consisting of:

1) linked area and list records (matched group)
2) the unlinked (residual) area records
3) and the unlinked (residual) list records.

We resolved all of the "possible match" ("clerical" in AUTOMATCH terminology) cases using the clerical review utility included with the software before the end of the linkage. Within the linked records, each group of records linked together consisted of one or more area records and one or more list records. There was no clerical review for the stage one links or the stage two, pass one and two links. We assumed that these were true matches. We conducted a clerical review for the stage two, pass three and four (individual name and address) possible matches, and these links were reclassified as matches or residuals. We did this review without regard to the overlap information contained on the files; that is, we did it as if we were using AUTOMATCH as an overlap tool before any other overlap detection procedure. The clerical review of possible matches took 4-6 hours to complete. With experience we were able to review about 200 pairs per hour.

After clerical review, seven possible outcomes of the linking process were considered:

1) AUTOMATCH overlapped the area tract to the same list record to which it was overlapped for the June survey. We considered these correct and did not review them further.

2) AUTOMATCH overlapped the area tract to a different list record(s) than the Wyoming SSO. While AUTOMATCH would still have overlapped the tract for the JAS, the possibility of incorrect overlap determinations for other surveys exists. These results were explored further.

3) One or more area records representing an overlap tract linked to one or more list records, but none of the list records was classified. Unless an unclassified record linked by AUTOMATCH is list-cross-referenced (LCR'ed) to the correct active record, then AUTOMATCH failed to correctly overlap this tract. These cases were explored further.

4) One or more area records representing an overlap tract failed to link to any list record. AUTOMATCH failed to correctly overlap this tract. We reviewed these cases to learn the reasons for the missed overlaps.

5) One or more area records representing a non-overlap or non-agricultural (non-ag) area tract linked to one or more classified list records. Either AUTOMATCH created a false link, or SSO personnel missed overlapping this tract. We reviewed these results to learn which was the case.

6) One or more area records representing a non-overlap or non-ag area tract linked to one or more unclassified list records. These records do not represent a problem, unless the unclassified list record was LCR'ed to another list record classified for the JAS. When this was the case, the tract was a missed overlap. We undertook a limited review of these cases to detect this sort of problem. JAS area tracts that linked to unclassified list records are the subject of a separate study to learn reasons for the list record's not being classified.

7) One or more area records representing a non-overlap or non-ag tract failed to link to any list record. We considered these results to be correct and did not subject them to further review.

Analysis of Link Outcomes--Table 1 summarizes the link outcomes before analysis. The tracts in categories one and seven were considered to be correctly handled by AUTOMATCH and were not subject to further analysis. Tracts in categories 2-6 were analyzed further. The analysis assumed that SSO personnel had correctly handled a tract unless there was obvious and overwhelming evidence to the contrary. Using this assumption, we were comfortable with the decision not to contact the SSO for help in the analysis. Empirical outcomes supported the assumption, since only a handful of unclear cases arose in which we would have questioned SSO personnel's decisions. The analysis also assumed that if AUTOMATCH linked an area record to a list record, and the list record contained a list-cross-reference to another list record which was classified and represented the same operation as the area record, this would be considered a success (that is, AUTOMATCH would have provided a path to the correct list record for overlapping the area record).

Outcome two contained 27 overlap tracts that linked to classified list records, but the list records were not the same ones to which the Wyoming SSO overlapped them.

1) In 21 cases, AUTOMATCH linked the area record to a list record LCR'ed to the correct overlap record. (AUTOMATCH succeeded.)

Table 1.--Wyoming Overlap Link Outcomes Before Analysis

Outcome	Number of Tracts
All outcomes	**1,661**[1]
Total JAS overlap tracts	**559**
1--Same as JAS	494
2--Link to different classified record	27
3--Link to unclassified record	17
4--Did not link	21
Total JAS nonoverlap and non-ag tracts	**1,102**
5--Link to classified record	34
6--Link to unclassified record	112
7--Did not link	956

[1]The total number of tracts did not agree between the JAS data file and the Area Name and Address Master (N&A) file. There were six tracts which appeared on the N&A file that were not on the JAS file and one tract on the JAS file that was not on the N&A file.

2) In one case, an operation was represented by two tracts; one tract contained one phone number for the operation, and the other tract contained a different phone number. The first tract linked to the list record for the operation on phone number; the list record was then not available to link to the second tract in subsequent passes. (This problem has been resolved in later versions of the software that allow the list frame records to be available to all passes despite having linked in an earlier pass.) (AUTOMATCH failed.)

3) In one case, SSO personnel had made a typographical error when entering the id of a list record in an area record's cross-reference field (the id entered differed from the correct id by one digit). (AUTOMATCH succeeded.)

4) In four cases, the area record linked to an incorrect list record that provided no path to the correct record. (AUTOMATCH failed.)

Outcome three contained 17 overlap tracts that linked to unclassified list records.

1) In nine cases, AUTOMATCH linked the area record to a list record LCR'ed to the correct overlap record. (AUTOMATCH succeeded.)

2) In four cases, the tract was cross-referenced to another state's list frame (the FIPS code for the other state was in the

cross-reference field of one of the tract's records). (AUTOMATCH had no chance to succeed or fail.)

3) In four cases, the area record linked to an incorrect list record that provided no path to the correct record. (AUTOMATCH failed.)

Outcome four contained 21 overlap tracts that AUTOMATCH did not link to any list frame record.

1) In four cases, the tract was cross-referenced to another state's list frame. (AUTOMATCH had no chance to succeed or fail.)

2) In four cases, including one extreme operator, the tract was not linked because the operation was represented by more than one tract, and another tract had linked in an earlier match pass on a variable not present in the tract in this category. The list records were, therefore, not available to link in later passes to these records. (AUTOMATCH failed.)

3) In four cases, including one extreme operator, the tract failed to link because the area and list records did not agree on the blocking variables. Since the records fell in different blocks, they were not compared and no link could be made. (AUTOMATCH failed.)

4) In four cases, tracts had names and addresses sufficiently different from the list records to which they were overlapped that AUTOMATCH would never link them. Two more tracts had only the same name in common with the list records to which they were overlapped (although, from geographical proximity of the place names, they appeared to be correct overlaps). (AUTOMATCH failed.)

5) In one case, the tract failed to link due to a name standardization error. (AUTOMATCH failed.)

6) In one case, the record was unavailable for evaluation because of a software bug (since fixed) (AUTOMATCH failed.)

7) In one case, the tract did not have a clear reason for failing to link. (AUTOMATCH failed.)

Outcome five contained 34 non-overlap or non-agricultural tracts that linked to classified list records. AUTOMATCH could not miss overlaps in this category or category six. Therefore, we considered that links that would be resolved on examination by SSO personnel as nonmatches were neither successes nor failures for AUTOMATCH. In other words, AUTOMATCH did not find any overlap in these cases, but it would not create an error, either.

1) In nine cases, the linked tract was an area frame record that was not available for estimation, as the name in the record had been changed. (AUTOMATCH neither failed nor succeeded.)

2) In eight cases, the tract was non-agricultural. (This is not too surprising since farms go in and out of business, and the resources are not available to send a criteria letter to every name on

the list frame every year.) (AUTOMATCH neither failed nor succeeded.)

3) In seven cases, AUTOMATCH made an incorrect link; mostly due to matching two records that erroneously contained the same phone number, or "close" matches (like a "John Doe Sr.'s" area record with "John Doe, Jr.'s" list record). (AUTOMATCH neither failed nor succeeded.)

4) In five cases, SSO personnel had missed overlaps. (AUTOMATCH succeeded.)

5) In two cases, a tract linked to both records of an "RS=85, RS=45" pair in which one record was classified and the other not. These operations represented by 85/45 pairs are treated separately on the list frame for sampling purposes, although they are operated by the same person. In both cases the classified operation was the one *not* represented by the area tract. (AUTOMATCH neither failed nor succeeded.)

6) In two cases, tracts representing the same operation were linked and cross-referenced to a classified list record. The tracts were correctly categorized as non-overlap, but the list record had been improperly updated in the SSO, creating an estimation problem (the operation was represented in the expansions of both the list and the area samples). (AUTOMATCH neither failed nor succeeded.)

7) One tract linked to two list records, one was an out-of-business 1987 list record LCR'ed to the other, a 1993 active, classified record. The only problem was that the area record contained the name on the 1987 out-of-business record as the operator's name. (AUTOMATCH neither failed nor succeeded.)

Outcome six contained 112 non-overlap records linked to unclassified list records. These cases have been analyzed as to the reason for the list record's not being classified in a paper by Orrin Musser [4]. (AUTOMATCH neither failed nor succeeded.)

Conclusions--Table 2 summarizes the performance of AUTOMATCH in this application. AUTOMATCH would be a useful tool for overlap checking. AUTOMATCH was successful, by the criteria of this analysis, in detecting 93.9 percent of the overlap detected in the state office. It detected 94.0 percent of the total overlap (the 559 overlap tracts found in the state office and the five additional overlap tracts found by AUTOMATCH). AUTOMATCH is not ready, however, to replace entirely the diligent efforts of an SSO employee in reviewing records. AUTOMATCH failed to overlap two extreme operators. One of these would be caught using version 3.0 of AUTOMATCH because of new capabilities, but missing any extreme operator is unacceptable. Further, AUTOMATCH needed review of its output to detect questionable links and to investigate link cross references to achieve these results. SSO employees did an exemplary job under pressure in detecting 99.1 percent of the total overlap. When AUTOMATCH becomes available in the state offices it should aid in the detection of

Table 2.--Performance of AUTOMATCH in the Wyoming Overlap Study

Outcome After Analysis	Tracts	Percent
Total Overlap Tracts	**559**	**100.0**
Tracts correctly overlapped by AUTOMATCH	525	93.9
Overlaps not detected by AUTOMATCH	26	4.7
Out-of-state overlaps[1]	8	1.4
Total Non-overlap and Non-Ag Tracts	**1,102**	**100.0**
Missed overlaps detected using AUTOMATCH	5	0.5
Other non-overlaps	1,097	99.5

[1]Since these require special handling anyway, they were separated here.

overlap under critical time pressures, but it is not a replacement for manual review, especially of extreme operators.

North Carolina New List Source Linkage

The single most important capability for a new record linkage system is the detection of matches between a new source list of farm operators and operations and a state list frame. By adding only nonmatches to the list frame, a list frame statistician can be confident that he or she is not adding duplicate records for a single operator or operation.

Purpose--The purpose of this linkage was to evaluate AUTOMATCH's performance in as rigorous a way as possible without having a pair of files for which we knew the "true" results for every record pair.

Matching Scheme--We linked the file of new source records to the active list frame name and address master records to find the records in the new source file representing operators and operations that were already represented on the list frame. The two files contained the following common identifiers:

1) name (one name on the new source, both primary and secondary names on the list frame),
2) street address,
3) place name,
4) state,
5) ZIP Code,
6) phone number.

Name and address standardization was carried out as follows:

1) Place names and ZIP Codes were verified and latitude and longitude were added using a C-language program developed in-house and a modified version of the NASS place name dictionary for North Carolina.

2) Addresses were standardized using the AUTOSTAN software package, along with a process developed by MatchWare for the Census Bureau's Geographic Division.

3) Primary and secondary names were standardized using AUTOSTAN along

Figure 2.--Matching Scheme for North Carolina New Source Match

Stage/Pass	Blocking Variables	Linking Variables
1/1	ZIP Code	Partnership Name
1/2	ZIP Code	"Corporation" Name[1]
1/3	ZIP Code	Primary and Secondary Name, Street Address, Place Name, State, Phone Number
1/4	Soundex Code of Primary Surname	Primary and Secondary Name, Street Address, Place Name, State, Phone Number
2/1	None	Phone Number[2]

[1]"Corporation" names are operation names other than partnership names, and include such name forms as "XYZ Farms" or "XYZ Ranches" as well as "XYZ, Inc." [2]This second stage was added after the results of the first stage were reviewed.

with a process developed in-house which utilized a modified version of the NASS name dictionary. All records contained standardized fields for corporation name, partnership name, and the components of individual names.

We adopted the matching scheme described in Figure 2. A few changes from the Wyoming Overlap Detection strategy are worth noting. First, we did not have SSN and EIN to link on since they were not present on the new source file. Second, we changed the way we used Phone Number. When we categorized the links in the Wyoming study, we discovered that phone number links were not very reliable. Because it is easily possible to get a false match on phone number, we decided that, from now on, all "phone number only" links must be reviewed. If we did a separate phone number pass before the name and address passes, we would have to review all of the record pairs with phone numbers that agreed. If we did the name and address passes first, we could assign many of these pairs to the linked subset based on name and address without review, thus reducing the number of record pairs to review manually. We first adopted a strategy of adding phone number to the name and address links, rather than making a separate pass. This did not work very well (see the Results subsection of this section). Adding a separate telephone number pass, in a separate stage, after the name and address passes worked much better.

Results of Linking--Linking divided the new source and list files into three groups, consisting of:

1) linked new source and list records,
2) the unlinked new source records, and
3) the unlinked list frame records.

Linking on the mainframe produced a similar set of outputs. Decisions made by record linkage operations (automated or manual) about record pairs fall into four categories. They are summarized in Figure 3. We evaluated the accuracy of AUTOMATCH as a matching tool by its ability to meet or exceed the accuracy levels of the current (mainframe) system as it is now used [5].

Figure 3.--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

We compared the results of the AUTOMATCH linkage to the results of the mainframe linkage. To make this task tractable, we assumed that for record pairs where the two systems agreed, the systems were correct. Our next step was to examine the lists of unlinked new source records from the two linkages. If the two linkages agreed, we again assumed that they were correct. If a new source record was linked to the list frame by AUTOMATCH, but not by the current system, we considered that this was a candidate for either a false match by AUTOMATCH or a false nonmatch by the current system. If a record was matched by the current system but not by AUTOMATCH, we considered that this was a candidate for a false nonmatch by AUTOMATCH or a false match by the mainframe. In the first case, we examined the AUTOMATCH link and made a decision about whether we believed it was correct. In the second case, we examined the mainframe link to see if we believed it was correct. Due to our inability to get the mainframe results in electronic format, this was an extremely time-consuming task, requiring that we examine over 3,000 mainframe links to see if they involved an AUTOMATCH nonlink or not. This process identified false links from the mainframe and false nonlinks from AUTOMATCH. For AUTOMATCH's errors, we attempted to learn the cause of the incorrect decision. There were 7,404 records in the new source file and approximately 77,000 active records in the North Carolina list frame file. The total number of true matches was 3,171. Table 3 summarizes our results.

Analysis of Link Outcomes--These results showed that, under the conditions of this test, AUTOMATCH clearly outperformed the current system as NASS uses it today. It is worth noting the consistency of the results between this linkage and the Wyoming overlap detection effort. In Wyoming, we concluded that AUTOMATCH missed approximately 6 percent of the overlap. This is not unreasonably different from the 4.9 percent false nonmatch rate of this linkage. Improvements in our method, on which we

Table 3.--North Carolina Match Error Rates--Current System vs. AUTOMATCH

	False Matches	False Nonmatches
AUTOMATCH	46 (1.1%)	155 (4.9%)
Current System	195 (4.6%)	275 (8.7%)

Table 4.--Causes of False Nonmatch in North Carolina New Source Match

Cause	Number of Cases	Per-centage
Multiple Causes	92	59.4
Address Differences	70	45.2
RR/Box # vs. Locatable (911) Address	23	14.8
RR/Box # vs. P. O. Box	4	2.6
P. O. Box vs. Locatable (911) Address	1	0.6
Address Standardization Errors	1	0.6
Missing or Different Individual Name	27	17.4
Different Partnership Name	14	9.0
Different Corporation Name	10	6.5
Different Phone Number	127	81.9
Phone Number Different, Rest of Records Agreed	18	11.6
Missing Phone Number[1]	28	18.1
Blocking Errors	2	1.3
Other, not specified	17	11.0
Total	**155**	**100.0**

[1]Missing phone number was not considered a cause of nonmatch for purposes of calculating "multiple causes," although disagreements in phone numbers were.

have already commented, may account for part of the difference. Also, a single operator on the list frame may be represented by more than one area tract. The area tract may, in turn, be represented by more than one record with different identifiers present in the different records. This complicated structure may have caused some false nonmatch. Of the 1,097 records that should not have linked in Wyoming, AUTOMATCH linked 15 of them, a false match rate of 1.4 percent. This agrees fairly well with the 1.1 percent false match rate in this linkage.

Even more interesting than comparing AUTOMATCH to the current system is an analysis of the causes of false nonmatch and false match in the AUTOMATCH linkage. Tables 4 and 5 summarize these causes. In Table 4, note that, because there were multiple causes of nonmatch in well over half of the cases, the numbers of cases for each cause do not sum to the total for nonmatches.

None of these results is unexpected or difficult to explain. It is not surprising that, in nearly 60 percent of the cases, it took multiple differences between the records to produce a nonmatch.

Because our linking strategy included a pass on telephone number alone at the end of the process, having only this variable correct was enough to produce a link. Thus, all of the false nonmatch cases either have phone number disagreements or missing phone numbers. For differences on this variable alone to produce a nonmatch, there had to be little discriminating power in the name and address information. (That is, very little information, very common values, or both.) This occurred about 10 percent of the time. In future linkages, we addressed errors in phone number creating or contributing to false nonmatch by eliminating phone number from the individual name and address linking passes. This had other advantages as well, which will be discussed below.

The second most common reason for false nonmatch was differences in address information. A locatable (so-called 911) address in one record and the rural route and box number style address in the other record caused about a third of these cases. This problem can only be expected to grow with the continuing conversion to locatable addresses. It is never legitimate in evaluating the performance of linking software to blame the quality of the data; if identifiers were complete and accurate on all records, then matching would be a trivial problem. However, we can improve matching performance by creating a strategy to handle the upcoming changes in rural addresses. We recommend that NASS acquire and carry both addresses on the list frame, at least for the near future.

It is also not surprising to find other address differences creating false nonmatches. A key assumption of the Fellegi-Sunter record linkage theory is the independence of linking variables. Empirical research at the Census Bureau [6] has shown that address information conforms least well to this assumption. The result is that address information receives agreement and disagreement weights that are too high relative to their true information content. This "overweights" disagreements in address information, causing the large contribution of address differences to false nonmatch. The strategy of removing phone number (and other numeric identifiers) from passes containing name and address has allowed us to adjust match and clerical cutoff weights to provide some relief from this problem. By assigning lower clerical cutoffs, we have allowed records with more disagreements on address variables to fall into the clerical review category.

Difference in individual (not operation or partnership) names is the next largest contributor to false nonmatch, being involved in a little over 17 percent of the cases. Again, the same weight cutoff adjustments that improve the address problems will relieve these problems, if they do not occur in combination with too many address differences.

Differences in operation (partnership or corporation) names contributed to approximately 15 percent of the false nonmatches. These problems were often due to our lack of experience with the string comparison metric function used by AUTOMATCH to compare alphabetic strings, and to the rudimentary

standardization applied to these names. The string comparison function computed a "distance" between the two strings based on the number of insertions, deletions, and transpositions required to turn one string into the other. This distance is expressed as a value which is used to prorate the match agreement weight for the two strings. Below some cutoff value (set by the user), the two strings are considered not to agree, and the match disagreement weight is assigned. In subsequent linkages, we experimented with a much "looser" agreement for multi-word character strings, and found that this reduced false nonmatches; however, it also increased false matches. The real solution to these problems came through better standardization. At the suggestion of the List Frame Section, we have developed more powerful standardization routines for partnership names that duplicate the methods used by the current system to break out both partners' names as individual names. This procedure would have eliminated many errors in the North Carolina study. Following discussions with Yue-Hwa Chang of the Bureau of Labor Statistics, who is using AUTOMATCH almost exclusively to link businesses, we adopted a new standardization approach for "corporate" names as well. The new process better uses the available information in the name while reducing false matches. Had this procedure been available, it would have prevented most of the "corporation" name false nonmatches in the North Carolina linkage as well.

While analysis of the false nonmatches reveals some problem areas, there were also two very positive results. First, address standardization seems not to be a problem, with only one case of false nonmatch due to a poorly standardized address. This augurs well for the quality of MatchWare's address standardization processes and AUTOSTAN.

Second, ZIP Code as a blocking variable and the use of multiple passes, with a final name and address pass blocked on the Soundex code [7], of the surname seems to work very well; blocking errors caused only two false nonmatches (1.3 percent). The use of the Soundex-blocked pass did not contribute a great deal to this result; eight additional false nonmatches (5.2 percent) would have occurred without it. While no records were kept regarding the causes of false nonmatch in the match using the current system, it was our subjective impression that using only one pass with the NYSIIS code as the primary blocking variable contributed significantly to false nonmatch. We found several cases where minor spelling errors were not overcome by the NYSIIS coding procedure and records were blocked apart. The current system does link across blocks on EIN, SSN, and phone number, but this output is not being reviewed operationally; therefore, the resulting pairs are treated as nonmatches. We suggest that, for maximum efficiency, NASS should always use multiple passes in future linking strategies.

While false nonmatch was a more serious problem than false match, and the consequences of false nonmatch are more serious in the construction of the list frame, we did evaluate false matches as well. Table 5 summarizes the characteristics of false match pairs.

In evaluating false matches, we tried to give the benefit of the doubt to the current system. If there was a reasonable argument that the two records might represent different operators, we designated the pair an

Table 5.--Characteristics of False Match Record Pairs in North Carolina New Source Match

Cause	Number of Cases	Per-centage
Different Operator with Similar Name	41	89.1
Parent/Child, Same Sex Sibling	34	73.9
Seniors vs. Juniors	5	10.9
Different Address	10	21.7
Different Phone	3	6.5
Different Address and Phone	9	19.6
Different First or Middle Name	7	15.2
Parent/Child, Spouse, Different Sex Sibling	5	10.9
Different Surname	2	4.3
Manager	1	2.2
Bad Phone Number Match	1	2.2
Individual Matched to Operation	2	4.3
Other	1	2.2
Total	**46**	**100.0**

AUTOMATCH false match. No assistance was requested from the North Carolina SSO. A similar linkage is being completed now in the Ohio Applications Research Section by Kara Broadbent [8]. In Ohio, SSO personnel are making decisions about record pairs on which AUTOMATCH and the current system disagree.

It is useful to remember the concept of discriminating power in examining the false matches. Almost 90% of the false matches were record pairs that had names that were sufficiently similar to link. These pairs had at least one other piece of information (a different initial, or slightly different address, or different phone number) that caused us to question whether the records represented the same operator. Nine of these cases, about 22 percent, had multiple disagreements.

Records with a different address were involved in approximately 40 percent of the false match cases. Usually an address disagreed on only one or two of the three to five components in a typical address. There was usually a positive agreement weight associated with the other components of the address. So, for all components considered together, there was often a positive agree-

ment weight. Again, overweighting of the address due to violation of the independence assumption is likely to be partially to blame. The same changes discussed in the section on false nonmatches improved this in later linkages. In addition, AUTOMATCH has an option that allows variables to be considered together (concatenated in AUTOMATCH terminology) when calculating match weights. We have not yet experimented with this feature, but we intend to do so.

We were pleasantly surprised by the low number of Sr./Jr. false matches; only about 10 percent of the false matches were of this type. We had expected this to be a bigger problem than it turned out to be. The other causes of false match were all negligible.

Conclusions--AUTOMATCH did well in detecting records on a new list source that represented operations already represented on the list frame. Compared to the current system, as NASS uses it, AUTOMATCH reduced false nonmatch by 43 percent and false match by 76 percent. Approaches developed in later linkages promise to improve on these results. The importance of this result is not to show that one system is superior to the other. Rather, it helps us conclude that AUTOMATCH is sufficiently accurate for NASS to use in a future record linkage system. The two systems are theoretically quite similar. We are certain that, if NASS used the current system as its designers intended, it would produce results much closer to AUTOMATCH's. In particular, adjusting match weight cutoffs and reviewing possible matches would bring the performance of the current system closer to that of AUTOMATCH. Such a review was envisioned by the designers of the current system as well, but is no longer done due to resource constraints. We feel strongly that provision of a possible match review capability that does not use too many resources is essential to the good performance of a record linkage system. We also feel certain that changes in technology will allow us to provide such a capability in any new record linkage system. (The total possible match review time for the AUTOMATCH was less than ten hours, due to its interactive review capability.)

Texas Unduplication

A close second in importance to detecting matches between a new list source and the list frame is the ability to detect matches within the list frame itself. Duplicates can appear on the frame for many reasons. One, obviously, is false nonmatch when adding a new list source. After a record is added, it may be changed because of criteria work or enumeration in a survey. The addition or correction of identifiers in the record may give us the information to link the record to its duplicate. Records may also be added to the list frame one at a time. Despite the best efforts, a record may occasionally be added that represents an operator or operation already on the list frame. With record linkage, we should find this duplicate.

Purpose--In June 1994, the Information Resources Management Committee asked us to use AUTOMATCH to estimate the amount of duplication within a list frame file. Texas was chosen to test the ability of AUTOMATCH to handle NASS's largest files.

Linking Scheme--To detect meaningful duplication (duplication which would affect estimates) on the list frame name and address master, we first attempted to extract active records, those that we believed to be subject

to sampling for NASS probability surveys. Due to a misunderstanding, we not only extracted these records, but some additional records as well. We will discuss this further in the analysis subsection below. We used AUTOMATCH's "undup" match type with our extracted file to detect duplication. The following identifiers were available on this file:

1) nine-digit record id,
2) SSN,
3) EIN,
4) telephone number,
5) primary name,
6) secondary name,
7) street address,
8) place name,
9) ZIP Code.

Name and address standardization was carried out as follows:

1) Place names and ZIP Codes were verified and latitude and longitude were added using a C-language program developed in-house and a modified version of the NASS place name dictionary for Texas.

2) Addresses were standardized using the AUTOSTAN software package, along with a process developed by Match-Ware for the Census Bureau's Geographic Division.

3) Primary and secondary names were standardized using AUTOSTAN along with patterns developed in-house which used a modified version of the NASS name dictionary. All records contained standardized fields for operation name (containing either the corporation name or partnership name), and the components of individual names.

We adopted the linking scheme described in Figure 4. There are some notable differences in this linking scheme from that for either of the two linkages discussed earlier in this paper. First, we did not use phone number with name and address or any other variable, but in two passes by itself. These passes were blocked--one on ZIP Code and one on the Soundex code of the primary surname. We parsed the phone number itself into three different variables, because we had noticed several records that looked like matches had phone numbers that disagreed only on area code. By separating the area code, we could make links on the remaining part of the phone number. We reviewed all phone number links. Operation names were no longer split into partnership and corporation name fields. Any operation name was standardized into a single operation name field. Passes one and two of stage two blocked these records on ZIP+4 Code and ZIP Code. We also included SSN, parsed into its three parts, to discourage links on similar names when SSN's were present and showed that the match would be false. Finally, two passes, blocked on ZIP+4 Code and ZIP Code, were made linking on name only.

Results of Linking--Linking divided the list frame records into two groups:

1) the unlinked, or unique records
2) the linked records, which represented potential duplication.

The same kinds of errors, false match and false nonmatch, occur in this type of matching as in the matches we discussed before.

Figure 4.--Matching Scheme for Texas Unduplication

Stage/Pass	Blocking Variables	Linking Variables
1/1	SSN	None
1/2	EIN	None
2/1	ZIP+4 Code	Primary and Secondary Operation Name
2/2	ZIP Code	Primary and Secondary Operation Name
2/3	ZIP Code	SSN (parsed into the first two, geographically-based regions and the last four random digits), Primary and Secondary Name (as an array), Address, Place Name, and State
2/4	ZIP Code	Phone Number (parsed into area code, exchange, and the last four random digits)
2/5	Soundex Code of Primary Surname	Phone Number (parsed into area code, exchange, and the last four random digits)
2/6	ZIP+4 Code	Name
2/7	ZIP Code	Name

Here a false nonmatch would be hard to detect. In the Wyoming overlap and North Carolina new source linkages, we had an independent source of information about the "true" matches. Unfortunately, no such information about the truth was available here. More, and presumably more accurate, information was available for linking list frame records in the unduplication case than in the new source linkage, and we had improved our methodology. There was every reason to believe that false nonmatch performance would be, if anything, better in the Texas unduplication than in the North Carolina linkage. Thus, we felt that the expenditure of months of time evaluating the false nonmatches here was an unwise use of resources. Detecting false matches is easier, since it requires only that an expert evaluate the linked pairs. The Texas list frame statistician agreed to do this; we present the results of his evaluation in Table 6.

About half of the linked records turned out to be nonduplicates. This does not mean that they were false matches. Often, they represented a match between a nonfarm operation and a farm operation or operator. This is not too surprising, since many farm operators may also operate or manage nonfarm facilities such as packing or canning plants, slaughterhouses, or grain elevators. Also, we relied solely on the record status indicators, but later learned that two other fields, agricultural data codes 29 and 30, which contained the farm code and *following year's* record status code, also affected whether the record would be subject to sampling in the following year (and, thus, of interest to us for duplicate detection). Had we understood their function, we could have used the agricultural data codes at the time we extracted records to remove nonfarm operations and operations which would not be subject to sampling in the next year.

Table 6.--Evaluation of AUTOMATCH Matches in the Texas Unduplication

Resolution of Matched Cases	Number of Cases	Per-centage
Nonduplicates	373	48.4
Duplicates	397	51.6
Agribusiness (Nonfarm)	202	26.2
Farms	195	25.3
Farm or Ranch Operation	78	10.1
Need Criteria Work to Resolve Status (Either Duplicate or Multiple Operations (Record Status Code Change to 85/45))	96	12.5
Multiple Operations (Record Status Code Change to 85/45)	21	2.7
Already Detected	14	1.8
Previously Undetected	7	0.9
Total	**770**	**100.0**

If we wished to detect duplication between nonfarm operations, we could include the farm code as a blocking factor in all stage two matches. AUTOMATCH's "select" feature could be used to segregate these records in the stage one links.

About half of the duplication we detected was in nonfarm records. Identifying this duplication was of use to the SSO, although it did not create a bias problem in NASS's agricultural surveys. The remaining 195 matches represented legitimate duplication of farm operator or operation records. Note that this represents only 0.18 percent of the total number of active records on the Texas list frame.

Of this total, 78 record pairs clearly represented duplication of farm or ranch operations. Another 21 pairs represented multiple operations owned by a single operator. These operations should have been coded using the special combination of record status codes 85 and 45, but had not been. Fourteen of these 21 record pairs had been detected by Texas SSO personnel before receiving our results. Another 96 record pairs required that further information be gathered, either by phone or mail (so-called "criteria work") before a determination could be made about whether they represented duplication of a single operation or should have their record status recoded using the 85/45 combination.

As a rough check on our assumption that we had not missed large amounts of duplication, we compared our results with the rates of duplication found by Musser and Mergerson in their studies of Ohio and North Carolina [9]. They listed their duplication rates by survey, but their median rates ranged be-

tween 0.42 percent and 0.18 percent. After analysis by the Texas list frame statistician, the overall rate of duplication we detected (for both farm and nonfarm records) was 0.36 percent.

Conclusion--AUTOMATCH is clearly up to the task of detecting duplication, even if we were not quite up to the task of specifying what records were to be checked. If the appropriate precautions are taken against including unwanted records and against linking nonagricultural operations and agricultural operators or operations, AUTOMATCH should be very successful in doing this task.

SUMMARIES OF OTHER MATCHES

Each of the three linkages described above explored AUTOMATCH's ability to do one of the three tasks that are critical for a new NASS record linkage system. We have also done some smaller linkages. These allowed us to learn more about AUTOMATCH or to show AUTOMATCH's ability to do a linkage task that would have been difficult to do with the current system.

U. S. Geological Survey Customer Service Survey

The main purpose of this linkage was to provide support for a reimbursable customer service survey for USGS in January 1994. As a part of the survey, USGS wished to create a sampling frame from lists of members of six mapping and satellite remote sensing organizations. Some of these lists were in ASCII format, others were in dBase files. After putting the files in a common format, we could go on with the linking.

First, the three files with usable address information were combined into a single file. We used AUTOMATCH's unduplication feature for the first time. With it we eliminated 439 duplicates. Next, we added the three files for which the address information was unusable for linking purposes. We unduplicated the resulting combined file on name alone, using two passes with the standardized name, one blocked on ZIP Code, the other blocked on Soundex code of the surname, and one pass with the unstandardized name. It was in this second linkage, with all of the files, that we encountered our only serious problem with AUTOMATCH in the two years we have used it. The clerical review software was dropping records. We went on by carefully updating some pointer files using a C program, and MatchWare corrected the problem immediately and sent us a new version. This second linkage detected 1,132 duplicate records, 35 of which came from the pass with the unstandardized names. (We would have benefitted in these cases from some improvements we've made since January 1994, in the name standardization process.) Overall, we detected 8.2 percent duplication across the files, a little below USGS's guess of 10 to 15 percent. The whole process, including cleaning up the input files (which were a mess), took about two weeks.

While the primary purpose of the linkage was to help the operational side of the agency, we did learn a few things. One was the importance of having a tool (or tools) available to manipulate data files. If files come in a standard format, this is not very important, but, if NASS uses record linkage for multiple purposes in the future, with files coming from many sources, it will be necessary to have such tools available. The second thing

we learned was the need to test new versions of the software thoroughly, and to pay attention to such things as record counts from pass to pass. Third, we learned that we needed better name standardization, which we have subsequently developed.

Farm Service Agency Racial/Ethnic Identifier Quality Assessment

Again, the main purpose of this linkage was to aid another project, this time one from the Survey Research Branch. There was interest in developing an ability to do rural surveys stratified by racial and ethnic identification of the respondent. One way to do this would be to add a racial/ethnic identifier to NASS's list frame. FSA collects such information when someone applies for FSA programs by having the FSA clerk code each applicant's race. This code is based on the opinion of the clerk, not a question asked directly of the applicant. By linking names of FSA applicants to NASS's list frame, a racial/ethnic identifier could be added to NASS's list frame for sampling purposes. We did this linkage to test the feasibility of using the FSA information.

In the 1994 June Agricultural Survey, NASS added a question about racial/ethnic identification to its area frame questionnaires. Using name and address information, and blocking on the first three digits of the ZIP Code, we linked an extract of the FSA data containing the racial/ethnic identifier to the June Area survey results. From the linked pairs we could extract a record that contained both the FSA and NASS June Area indicators of race. Our results failed to show that the FSA indicator of racial/ethnic identification was sufficiently accurate for sampling purposes. Interested readers may contact Mike Fleming of the Survey Quality Research Section for more detailed information.

Again, we did not learn a great deal that was new about AUTOMATCH from this match. Mainly, we showed AUTOMATCH's flexibility and the usefulness of having generalized matching software available that can be turned to many tasks, rather than a purely proprietary system that is specialized for list frame maintenance.

Washington Cattlemen's Association

The purpose of this linkage was to help the Washington SSO with a reimbursable survey for their state's cattlemen's association. The Washington SSO requested our support because of AUTOMATCH's ability to produce many different output files from a match and the List Frame Section's busy schedule. The cattlemen's association wanted to survey all operations with cattle in the state, and was willing to give the SSO the cattlemen's association list for list frame maintenance. The SSO wanted two different outputs from the linkage. The first was a list of all of the cattle operators on the cattlemen's association list that were not on the Washington SSO's list frame, for purposes of adding these records to the list frame. The second was a list of their list frame operators and operations that did not appear on the cattlemen's association list. They planned to use this list, with the cattlemen's list, to do the survey.

The format of the cattlemen's list allowed us to try out our AUTOSTAN process for "surname on the left" names. We also used a modified version of our regular name standardization process on the secondary name field, which often consisted simply of a first name if the surname was the same as

the primary name. Also, we used AUTOSTAN to filter out some "junk" (non-name information) about membership status. We used AUTOMATCH's array matching feature to match the two phone numbers on the cattlemen's list to the one phone number on the list frame. We were surprised that we did not get more links on the work phone. While no rigorous evaluation of the linkage was done, the Washington SSO was pleased with AUTOMATCH's performance in detecting duplication.

Again, this linkage made clear to us the desirability of using AUTOMATCH and AUTOSTAN to their full capabilities when facing a nonstandard list source. Also, we saw the value of experience in deciding how to treat the various fields. For example, we combined a separate last name field and first name field simply by defining them as one field and standardizing them using the surname on the left patterns.

Agricultural Marketing Service Dairy Lists

In December 1994, Systems Services Branch (SSB) approached us for help in linking the 1993 and 1994 AMS Dairy Lists. Before that time, in each state, each year's AMS list was presented to the list frame statistician as a potential source of new dairy operators. Unfortunately, many names contained on this list are the same from year to year, and are therefore of little use. It was proposed that we link the 1993 list to the 1994 list and extract the 1994 nonmatches to make a useful list of new dairy farmers.

Because, in theory, names that appeared on both lists should not have been changed, we first attempted a "quick and dirty" character for character link on name between the two files. This yielded fairly good results (and, frankly, could have been done just as easily without any sophisticated linking software); however, we were still able to find some records for which name or address corrections had been made which caused the character for character link to fail. Therefore, we proposed to teach SSB staff to use AUTOSTAN and AUTOMATCH, to make more sophisticated matches. After two days of training, they improved on the simple match and eliminated many duplicates. They were still somewhat disappointed with the results, since they wanted to eliminate *all* of the duplicates.

They also felt that AUTOMATCH was difficult to learn. Members of the List Frame Section, whom we had also taught to use the software, and who are familiar with the current system and record linkage, concurred. Among those trained to use AUTOMATCH, only one person has not expressed this opinion. Perhaps this is because she received three and a half days of training, and was committed to becoming an expert in the use of the software as a primary part of her project responsibilities. The others have many competing responsibilities, and do not feel they have time to become "experts." This is likely to be the case with most state list frame statisticians.

This was the first linkage in which we used AUTOSTAN to generate individual names for partners in a partnership, allowing partnership records to link on name to records representing one partner. We were pleased with the results it gave. Also, MatchWare had developed and provided a new address standardization process that produced a more efficient and smaller standardized address. We were equally pleased with the results of this new process.

The main thing we learned in this linkage was that AUTOMATCH, without any additional "front ends" to simplify the linking strategy and parameter file creation processes, is perceived as difficult to learn and use. Systems Services Branch is now working with us to create front end software to simplify the use of AUTOMATCH.

California Fruit Chemical Use Survey

The most recent linkage we have undertaken, in August 1995, is a linkage between a file of California EPA permit holders and the 1995 Fruit Chemical Use sample. The California EPA issues permits to use pesticides. They maintain two databases: one of permit holders, with their names, addresses, phone numbers, and permit numbers; and a second file of permit numbers and the amounts and types of pesticides applied. To use this administrative information to estimate levels of pesticide use in NASS's fruit chemical use program, it is necessary to associate a permit number with each operator in the sample file who holds a permit. Simply asking respondents for their numbers results in some inaccurate numbers being collected and in nonresponses. The Survey Quality Research Section suggested that AUTOMATCH might be an appropriate tool for this match.

We used yet another new version of our name standardization process to standardize the names on the EPA list and on the fruit sample. This version separated each corporation name into a unique part and up to three keywords (like Inc., or Ranch). We also used MatchWare's most recent address standardization process. In addition, this linkage gave us a chance to test version 3.0 of AUTOMATCH, which incorporated some features, such as zero weights for missing values, for which we had been waiting.

We observed encouraging results with the new standardization software and the new features of version 3.0 of AUTOMATCH. A preliminary link on California EPA ids which were already in the sample file from previous years' efforts yielded 600 links. Linking on name and address resulted in another 1,000 or so links. The SSO chose other methods for accomplishing this task and did not use the AUTOMATCH results.

CONCLUSIONS

AUTOMATCH can do all of the tasks necessary to be the core component of a new NASS record linkage system. We tested AUTOMATCH's ability to link a new source to a state list frame in the North Carolina linkage. AUTOMATCH outperformed the current system as it is currently used. We anticipate that, if we repeated this work using the methods we have since developed, we would improve on this performance.

We tested AUTOMATCH's ability to unduplicate an existing list frame, and to handle NASS's largest list frame files, in the Texas unduplication. The results we got were consistent with earlier research estimates of list frame duplication. AUTOMATCH had no trouble handling the Texas list frame file.

Linkages similar to the North Carolina new source and Texas unduplication have been completed or are under way in the Ohio SSO under the direction of Kara Broadbent. She will prepare a separate report when they are complete. Based on preliminary results, we expect these linkages to provide independent confirmation of the results we observed. We also expect that this work will produce a pair of files with known results, which NASS can use in future research work.

Another record linkage task is detection of overlap between the area frame sample and the list frame in the June Agricultural Survey. The Wyoming linkage tested AUTOMATCH's ability to accomplish this task. We consider this effort a qualified success. AUTOMATCH did not do as well as human experts, but its use could significantly reduce the time and resources necessary to do the task during June pre-survey period.

NASS also has other uses for record linkage that are not easily met with the current system. This includes building frames for reimbursable surveys and adding administrative data to survey files. In the smaller linkages we reported on, AUTOMATCH showed that it can add these new capabilities.

During our two-plus years of experience with AUTOMATCH, we have found only two areas of real concern. When we first began using AUTOMATCH, it was relatively new on the market (only three years old) and MatchWare was just beginning to achieve some real commercial success with it. We encountered several minor bugs, and one major problem. While MatchWare responded quickly to our reports of problems, we worried that each new version seemed to require testing before we could trust it. Over time, however, this problem seems to have disappeared. We believe that this is due to a change in philosophy by MatchWare. When MatchWare first released AUTOMATCH, a small cadre of people with expertise in both computers and record linkage made up most of the users. Matt Jaro, the president of MatchWare, said these early adopters tended to have a "pioneer mentality"; they wanted new features and better performance as rapidly as possible, and were not concerned about the occasional need to work around a bug. Now that the product is mature and being purchased by a wider range of users, he says that his focus has shifted to providing a stabler, more reliable product, with much more testing done before the release of new versions. The time between new versions has increased, as has their reliability. We have detected no problems in testing version 3.0, confirming what Jaro told us.

The second problem we have is with perceived ease of use. While we believe that record linkage is not a trivial task, and that no record linkage software would be much easier to use than AUTOMATCH, we still acknowledge the validity of this concern. If we expect list frame statisticians in individual states to use the new system, we must simplify the system further and provide adequate training as well. Particularly in smaller state offices, there is no time for state office personnel to become experts on yet another software package, and the theory and practice of record linkage. To this end, we are now engaged in a joint project with Systems Services Branch to develop interfaces that will simplify the coding of parameter files. In addition, default parameters for common applications are being developed.

RECOMMENDATIONS

We recommend that NASS obtain and use AUTOMATCH and AUTOSTAN as the basic component of the record linkage/resolution sub-systems of ELMO. While AUTOMATCH does record linkage using the Fellegi-Sunter method, and can be used temporarily as it comes from MatchWare, there are several enhancements necessary for it to meet all of NASS's functional requirements. We further recommend that NASS develop these enhancements. We have divided the tasks necessary to

complete these enhancements into systems development tasks, additional research, and implementation recommendations. The additional research and systems development tasks are similar to recommendations adopted by the ELMO2 steering committee in their working paper on record linkage.

There are three alternatives to using AUTOMATCH. The first is adoption of another commercially available record linkage solution. We reviewed the available software in *Record Linkage I: Evaluation of Commercially Available Record Linkage Software for use in NASS*, and concluded that AUTOMATCH was the best choice. Nothing in our evaluation of the product has changed our minds.

The second alternative is porting the current system of legacy COBOL and FORTRAN programs to a new platform. This system of programs is no longer well understood. It would take considerable effort simply to document the functions of the different programs that make up the system. To this would be added the effort of modifying and recompiling the programs. Significant modifications would be required to match AUTOMATCH's functionality, flexibility, and efficiency. While we do not doubt NASS's ability to accomplish this task if it were necessary, we are certain that the resources involved would cost many times the price of AUTOMATCH. In addition, all of the programming of interfaces required for building a new record linkage system using AUTOMATCH would still be required using the current system as a core component.

The third alternative is to build a new record linkage system from "scratch." This might be worth considering if there were methodological problems with AUTOMATCH or the current system, but both AUTOMATCH and the current system use the most widely accepted methodology for doing automated record linkage. To build a new, proprietary system would require thousands of hours of development effort. This would, obviously, be many times more costly than simply purchasing and using AUTOMATCH as the core of the new record linkage solution.

Systems Development Tasks

Provide an interactive interface for the possible match resolution part of the system.-- Statisticians need to review "potential matches" that the system cannot resolve. We feel that the largest part of the performance difference we observed between AUTOMATCH and the current system in North Carolina was due to our review of potential matches from AUTOMATCH. Both in theory and in practice, reviewing more records results in lower error rates, and decreasing returns result from additional review. This implies that allocating at least a small amount of time to a review of possible matches may give a large benefit.

Our experience with batch systems for this activity is that the time, effort, and volume of paper involved in this manual process are prohibitive. Therefore, SSO's typically let the system do what it can and get to the resolution output only as time and resources allow. AUTOMATCH has a rudimentary online review capability, but it lacks some critical functionality. We need an online review capability that will allow statisticians to efficiently review output and resolve potential matches. This system should allow for multiple simultaneous reviewers and the capability of writing notes "on the fly." The system should monitor and record the attempts to resolve potential matches. In addition, it should have links to

the database that allow it to extract comment fields for records being reviewed and display them as hypertext. It should also produce paper output if requested. This system should have online help available.

Create a system for place name verification.--NASS needs to create a process, either using Sybase or a separate program, to standardize place names and validate zip codes, and generate a latitude and longitude value for each record for use as part of a distance function for linking. (The place name based latitude and longitude may eventually be replaced by Global Positioning System coordinates.)

Build a graphical front end for AUTOSTAN and AUTOMATCH.--AUTOMATCH and AUTOSTAN are powerful, flexible programs, but using all of their capabilities comes at the price of coding each parameter to specify a linkage in the appropriate parameter file. This tedious and time-consuming process requires a knowledge of both AUTOMATCH and record linkage. NASS needs to develop a front end that automates this process, and gives the less expert user access to several initial setups that an SSO list frame statistician can use for routine applications. In addition, the front end should simplify the process of creating "custom" linking solutions. Such a front end should have a graphical user interface so that users can specify a linkage by making choices from a menu. The front end programs could then create the needed parameter files and launch the necessary AUTOMATCH and AUTOSTAN applications to do the linkage. The user interface should handle all parameter file generation, including the files that specify extracts and reports, and those that specify input file formats. In addition, any front end tasks, such as assigning list ids to the incoming data records, should be handled by this program. This software should have online help available to guide the user.

Enhance the Capability to Store Information About Previous Linkages.--AUTOMATCH can carry linkage information between several passes of data within the same linkage, but NASS needs to develop the capability to store that information for a longer period of time across linkages. For example, a state may link an FSA data file to the list frame every year. A new record linkage system needs to provide the capability to detect linkages between identical record pairs from year to year to avoid having SSO personnel resolve the same possible matches year after year. One objection to using the current system as designed has been precisely this kind of repeated effort. If we cannot eliminate this repetition of work, we run the risk of seeing the same kind of short cuts applied to the new system that are being applied to the current system.

Enhance Extract Capabilities.--NASS needs the capability to add complete records to the list frame based on results from the record linkage and possible match resolution processes Software needs to be created that takes AUTOMATCH flat file extracts, formats them properly, and adds the resulting data to ELMO. This software should build a single record from several matched and duplicate records to capture any new information that might be added from a linkage, and add new records based on unlinked new source records. There needs to be a menu driven interface for users to do these tasks.

The ELMO database contains separate tables for Person, Operations, and Person/Oper connections. Record linkage, possible match resolution, and the procedures to add whole or partial records to the data base must recognize

and use this system of tables. In particular, new records must be generated so that all appropriate tables are updated.

Allow real-time duplication checking.--This would be useful, but it is lower on our priority list of enhancements because the "look-up" processes in ELMO can do the same function. There is considerable variability between users on how intensively they search for duplication before adding a new record. This feature would define and enforce a minimum level of duplication searching and do it automatically before a record is added. AUTOMATCH and AUTOSTAN Windows DLL's and UNIX callable libraries exist for creating interactive duplicate checking systems.

Additional Research

Support Systems Services Branch.--The Research Division's priority during the implementation period for ELMO is to support Systems Services Branch's efforts to do the tasks listed under "Systems Development Tasks."

Develop efficient "weight cutoffs" for matching algorithms.--AUTOMATCH requires the user to set weight cutoffs. As part of the new front end, we need to be able to automatically assign reasonable efficient weight cutoffs based on the number and types of variables included in the match. We need to develop a utility to do this for inexperienced users.

Create ELMO extract specifications.--To use the batch (non-interactive) versions of AUTOMATCH, an extract of name and address records from the list frame has to be created. AUTOMATCH will use a standard extract each time it is used with the list frame files. Specifications for creating this extract from ELMO need to be written.

Create standard parameter files for linking FSA files to the list frame, for doing unduplications, and for overlap detection.--Research Division needs to create a standard set of parameter files for linking FSA lists (our most common new source) to the list frame, for unduplicating the list frame, and for detecting overlap between the list and area frames in multiple frame surveys. As we envision it, these kinds of matches should be specified by supplying one or two file names and clicking a button. (If FSA record formats change, these files can be updated.)

Create an AUTOSTAN process for doing postal address standardization.--In order to receive postal service discounts, NASS needs to standardize addresses to meet postal service specifications. Research Division should explore AUTOSTAN's capabilities to do this task.

Improve the AUTOSTAN process for doing name standardization.--There are still some name forms, in particular FSA names of the form "Smith, John jt vent," or "Jones/Smith/Jackson j.v.," which AUTOSTAN is not equipped to handle. Research Division should modify the name standardization process to handle these name forms.

Implementation Recommendations

Simplicity and ease of use.--It has been argued, with considerable justification, that any new linking system needs to be very simple to use. Our agreement with this statement is qualified. As far as it is possible, we should simplify and speed the matching process through the creation of good front and back end interfaces. We will develop a standard procedure for some frequently done linkages, such as linking FSA files to the list frame or unduplicating a

list frame. Using such a procedure, a novice could obtain good results with a minimum of instruction.

However, for nonstandard list sources, obtaining good matching results requires the application of some expertise with both record linkage and AUTOMATCH. Therefore, *we recommend that a group of experienced list frame statisticians be given additional training so that they can help less experienced list frame statisticians from other states when they encounter unusual list sources or have specialized uses for matching*. Alternatively, such expertise might be located in the List Frame Section in Headquarters.

We do not recommend, under any circumstances, dispensing with the review of possible match cases.--Along with others participating in the creation of a new record linkage system, we are obliged to create a possible match review capability that provides all the needed functionality, is as easy to use as possible, and allows the user to resolve possible matches in as little time as possible. Development of such a utility is the appropriate response to limited time and resources for possible match review, not elimination of this crucial step in the linking process.

This opinion is not only our own, but was expressed at the time of the creation of the current NASS system. ". . . it is necessary not only to stress the amount of time and effort that will be required in the manual resolution process, but also the importance of this stage. . . . It is the resolution phase that actually allows review of the automated resolution procedure and to fully resolve duplication. The job performed by each field office during resolution will determine the resulting quality of the list sampling frame" [10].

Unfortunately, the technology available then did not allow the creation of a system that supported manual resolution of possible matches with an acceptable level of resource expenditure. We believe the technology now exists to realize this long-standing goal.

Monitoring system functioning and continuous improvement.--Everyone involved in the process of creating the new record linkage system will try to make it as good as possible when it is implemented. Nevertheless, it is necessary that someone monitor the functioning of the system, particularly the functioning of the default parameters. Over time, these parameters may need adjustment to provide the best possible performance.

Further, efforts should be made, as resources are available, to continue to improve the performance of the system by monitoring problems that arise and attempting to adjust parameters to eliminate them. This may be especially important with AUTOSTAN. As more names and addresses are processed and new patterns of errors in standardization emerge, classification and pattern files can be modified to allow more name and address forms to be standardized correctly.

Carry two addresses on the list frame during the transition period to locatable addresses.--It may be some time before all of the addresses on our list sources are updated to the new locatable (911) address. Carrying both the rural route/box number and locatable addresses will increase the number of correct matches by providing additional linking information.

Find or develop standard tools for manipulating input data files.--We anticipate no problems extracting the list frame records needed for record linkage from ELMO in the desired format using ELMO's powerful tools for this purpose. However, we have found it the rule rather than the exception that some sort of manipulation is required for new source files. We have often relied on C programs for this purpose, because we were familiar with C and C is very powerful. We were impressed, however, with the ease with which we could accomplish the needed manipulations (such as creating fixed field and record lengths, or concatenating nonadjacent variables) with dBase in the USGS project. We would recommend that a similar tool be adopted for use as a standard way of manipulating input files for AUTOMATCH.

Use multiple matching passes.--It was clear from our results that the use of multiple matching passes decreased both types of linking errors. We recommend that multiple passes always be used.

Use AUTOMATCH as an aid to overlap detection, but not a replacement for human review.--The Wyoming project convinced us that AUTOMATCH could detect almost all of the overlap between list and area frames in a multiple frame survey. However, its failure to detect an extreme operator leads us to make the qualified recommendation that AUTOMATCH be used only as an aid in overlap detection.

This report's recommendation of AUTOMATCH for NASS's use is based on existing and planned applications and is not in any manner a general recommendation on record linkage software outside NASS.

REFERENCES

[1] Musser, Orrin, and Mergerson, James W., *Estimating List Frame Duplication*, Survey Research Branch, Research Division, National Agricultural Statistics Service, USDA, 1994.

[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] Broadbent, Kara, *Record Linkage III: Experience Using AUTOMATCH in a State Office Setting*, Survey Research Branch, Research Division, National Agricultural Statistics Service, USDA, forthcoming.

[4] Musser, Orrin, *Analysis of List Coverage Problems and the NOL Domain*, Survey Research Branch, Research Division, National Agricultural Statistics Service, USDA, 1995.

[5] The current system was designed for use in stages, with corrections made to records that could not automatically be standardized, and possible matches reviewed. Because of the technology available in the late 1970's, these functions were very time consuming. It became clear when the system was implemented that the resources to follow the steps envisioned by the developers were not available, so the review steps were dropped in favor of posting unlinked records to the list frame as inactives, subject to criteria work.

[6] Winkler, William, "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 1988, pp. 667-671.

[7] The NYSIIS phonetic code was not available in AUTOSTAN before version 3.0 was released. NYSIIS is now available, and will be used in the future in place of Soundex.

[8] Broadbent, Kara, *op. cit.*

[9] Musser, Orrin, and Mergerson, James W., *op. cit.*

[10] Arends, William, *Methodology for the Development, Management, and Use of a General Purpose List Sampling Frame,* Statistical Reporting Service, 1977.