THE SET THEORY MATCHING SYSTEM

Bradley E. Slaven, Bureau of the Census
Bureau of the Census, CSMR, Washington, DC 20233-4700

Keywords: Record Linkage, Administrative Record Matching System

1. INTRODUCTION

The Set Theory Matching System is a computer system application which is used to match data records from two independently collected data sources. The two data sources used in this study are the 1990 Census and the 1990 Alternative Enumeration.

The research project this system was designed to assist is an effort to identify in detail the characteristics of people who are missed in a standard census enumeration, and the circumstances which lead to less than complete enumeration. The Set Theory Matching System aids this objective of the project by identifying those individuals and households that have been enumerated in both data sources, only in the 1990 Census data source (i.e., possible false enumerations), or only in the 1990 Alternative Enumeration (i.e., possible missed enumerations).

This paper discusses the theoretical applications of the hierarchical decision theory involved in the Set Theory Matching System and presents results evaluating the statistical accuracy of the system as measured by field research.

The matching process is complicated by the occurrence of inaccurate, incomplete, and conflicting data in the two data sources. The housing unit information used by the system includes street name, street number, and a unit designator (e.g., apartment number). The occupants' names and several demographic characteristics--age, year of birth, marital status, sex, and race--are also included.

The Set Matching System attempts to resolve the data differences by a complex set of algorithmic decisions. The system uses a hierarchical decision system which considers whole households as the primary unit of matching consideration and the evaluation of household members as the secondary matching criterion.

The system groups the Census and Alternative Enumeration data into sets of households. The household sets are in turn composed of sets of household occupants. By using whole households as the primary unit of matching consideration, the demographic information on multiple household members becomes the guide, rather than simply attempting to match one individual from one data source to another person in a second data source. The matching process can exploit the demographic information of the whole household and is less dependent on the potentially inaccurate demographic information of a single individual. Therefore, increased match rates and decreased error rates can be expected from this process.

2. THE DATA SOURCES

Two independently collected data sources are used in this research project, the 1990 Census and 1990 Alternative Enumeration. The data sets are independently collected censuses of the same geographical locations at approximately the same time period. The Alternative Enumeration data were collected as part of a large scale ethnographic enumeration research project in 28 nationwide sites. The research sites were all located in areas where severe underenumeration in the 1990 Census was expected. The Census data were collected in standard fashion, either by mail response to the census, or by U.S. Bureau of the Census enumerators during the 1990 follow-up enumeration.

3. THE IMPORTANCE OF USING HOUSEHOLD INFORMATION IN THE RECORD MATCHING PROCESS

Traditional methods of matching individual data records from two data sources compare each individual record from one data source to the data records of a second source. Many highly sophisticated matching process algorithms have been developed which link individual records from two data sources, but which ignore household occupant related information (Fellegi and Sunter, 1969; Winkler, 1985, 1987a, 1987b; Statistics Canada 1982, 1984). When household occupant related information is incorporated into the matching process, the demographic information concerning all the occupants of the entire household can be utilized. Using individual records to link people presents a problem if, for example, in one data set the wife is listed by her maiden name, while in the other she shares her husband's surname. By inspecting the entire household, the fact that the two records match is much more obvious, since the match is less dependent on the potentially conflicting, inaccurate, or missing demographic information of a single household occupant. The availability of a greater amount of data regarding the entire household may well yield increased match rates and decreased error rates.

The more traditional record matching systems have not exploited use of household occupant related information for two basic reasons. First, most record matching systems were developed prior to the advent of modern programming languages, such as Pascal, C, and PL/I. Incorporating household-related information into the record linkage process requires the flexibility of such languages. Unlike traditional programming languages such as COBOL, FORTRAN, and BASIC, modern languages allow the software engineer to allocate computer memory dynamically through the use of pointers and create a wide range of abstract data structures that can be implemented as linked lists. The additional flexibility of a modern programming language allows the software engineer the opportunity to organize the data in a hierarchical form that better represents the natural order of the data and preserves the unit of population enumeration. The advantage of aligning the natural order of the data to computer application development cannot be overstated. Data structure creation is often the fundamental and single most important element in any software development effort (Kernighan and Ritchie, 1988; Wirth, 1976).

Secondly, the use of household-related information in the Set Theory Matching System requires multiple passes through the data sets as opposed to more traditional single-pass record matching systems. Multiple passes require additional computer processing time, although, as in the present system, this limitation can be algorithmically reduced to a considerable degree (Horowitz and Sahni, 1987). In addition, most computer matching systems include a clerical review process to examine the results of the computer matching system. The time required for the clerical review process can be greatly reduced with the multi-pass hierarchical matching system because of increased match rates and decreased error rates.

4. MATHEMATICAL SET ORGANIZATION OF THE DATA RECORDS

The organization of the Census and Alternative Enumeration data sets is fundamental to understanding the theoretical application of the Set Theory Matching System. The Set Theory Matching System views the Census (S_1) and Alternative Enumeration (S_2) data sets as hierarchically organized mathematical sets. Household elements are the primary unit of matching consideration and contain housing unit identification information. Within household elements, person elements are the secondary matching unit and contain demographic information concerning each household occupant.

Mathematically, these hierarchical sets are described as follows: S_1 and S_2 are the Census and Alternative Enumeration data sets and H_{ij} represents a household set such that

$$H_{ij} \in S_i$$

for $i = 1$ to 2 and $j = 1$ to n_i, where n_i is the number of households in data source i.

In addition, H_{ij} is a set of household elements such that

$$P_{ijk} \in H_{ij}$$

for $i = 1$ to 2 and $j = 1$ to n_i, and $k = 1$ to m_{ij}, where m_{ij} is the number of persons within household H_{ij}.

5. OVERVIEW OF THE SET THEORY MATCHING SYSTEM

The Set Theory Matching System uses four stages of matching algorithms, primarily based on household membership, to perform the matching procedure. The process is based on mathematical set theory. As households are linked and housing unit occupants are matched, the elements are removed from further consideration. This method allows for a reduction in the linking and matching requirements without a reduction in the accuracy of the matching process.

The first two stages of the procedure attempt to link households from the two data sources. The third stage matches the household occupants from the previously linked households. The last stage of the procedure attempts to find additional matches between the people of the data sets regardless of household affiliations.

The Set Theory Matching System was developed using test data from a prototype Alternative Enumeration conducted in conjunction with a census test in 1988. It was expected that the prototype Alternative Enumeration and pre-census data would be similar to the 1990 Census and Alternative Enumeration data sets. Heuristic matching rules were developed empirically using the 1988 data in anticipation of the 1990 Census and Alternative Enumeration.

The Set Theory Matching System was intended to maximize the accuracy of the record matching process. However, the system assumes that Type I ("false positive") errors[1] are a more serious problem than the Type II (false negative") errors[2]. The Type I errors are a more serious because the errors produce incorrect matches, as opposed to the Type II error of overlooking a true match. Since all incongruities are reviewed and resolved, it was deemed better to fail to assign a "matched" status, because the error was more likely to be detected and corrected in the field research stages of the project. Type I errors, on the other hand, would appear to be congruent enumerations and thus would tend not to receive sufficient attention for error detection.

5.1 STAGE ONE: THE PRIMARY HOUSEHOLD LINKING PROCESS

The primary household linking process consists of an iterative series of algorithms that evaluate all potential household matches with each iterative pass. The iteration process begins with strict household linking requirements which are gradually relaxed with continued passes of the data. Initially, each household element from the Census (S_1) set is compared to each household element in the Alternative Enumeration (S_2) set. That is,

$$\forall H_{1j} \in S_1$$

and

$$\forall H_{2j} \in S_2$$

A link exists between a household in S_1 and a household in S_2 if either a minimum of: 1) R_q percent of the persons in a household (H_{1l}) are mapped to persons in a household (H_{2m}) for some $l <= n_1$ and some $m <= n_2$; or 2) R_q percent of the persons in a household (H_{2m}) are mapped to persons in a household (H_{1l}) for some $l <= n_1$ and some $m <= n_2$; where R_q decreases from 1.00 toward 0.00 as the iteration number (q) increases.

Although the person mapping process ultimately serves the purpose of matching individual records, in the initial matching stage it is simply a means of linking households between the two data sets; no actual person-level match is attempted at this stage for the occupants of each household. The requirements for mapping a person from H_{1l} to H_{2m} primarily take demographic characteristics, first name, and last name into consideration. The mapping process uses an extension of the Soundex Algorithm to evaluate the name fields. An age and year of birth sensitivity indicator, similar to an age range, is used to evaluate whether potential occupant matches are of similar age. (The extended Soundex Algorithm and age and year of birth sensitivity indicator are discussed later in the paper.)

When the Set Theory Matching System identifies a matched household pair, those household elements are put together and removed from further household linking consideration. Therefore, each pair of matched household elements (H_{1l} and H_{2m}) redefines the subsequent Census and Alternative Enumeration matching universes. That is, $S_{1q} = S_1 - H_{1l}$ and $S_{2q} = S_2 - H_{2m}$, where l represents the index of a linked household in S_1, m represents the index of the corresponding linked household in S_2, and q is the iteration number of the matching process.

As the iterative matching process continues, the number of unlinked household elements remaining in the sets (S_{1q} and S_{2q}) decreases. This reduction in the number of remaining potential household links justifies a reduction in the algorithmic requirements necessary for linking subsequent household elements, because the remaining unmatched households become increasingly diverse. Therefore, the percentage of persons required for a link (R_q) between households decreases with continued iterations.

5.2 STAGE TWO: THE SECONDARY HOUSEHOLD LINKING PROCESS

The secondary household linking process provides a means to evaluate the household linking status when limited household occupant information is available. For example, vacant housing units can be matched and referenced across the data sets by this process. In addition, this stage allows for the possibility that a different set of people could be enumerated at a particular housing unit location due to occupant mobility.

The secondary household linking process uses two iterations of housing unit address information to attempt to link additional households from S_{1q} to S_{2q}. The first iteration uses the street number, street name and unit designator fields to attempt to link household addresses. As is the case in stage one of the household linking process, when household addresses are linked together they are removed from further matching consideration. Therefore, the pool of eligible household links in the sets S_{1q} and S_{2q} continue to decline throughout the entire linking process.

The last iteration of the secondary household matching process uses the street number and street name to match households. The unit designator information is not used. Linked households are removed from further household match consideration.

The housing units that remain unlinked throughout the household linking process are not processed through stage three, the primary person-level match using matched households. However, the occupants of unlinked households still receive occupant matching consideration in the outside the household individual matching process (stage 4).

5.3 STAGE THREE: THE LINKED HOUSEHOLD OCCUPANT MATCHING PROCESS

The household occupant matching process is an iterative routine that reviews all the linked pairs of households and matches the corresponding household occupants. Mathematically, that is

$$\forall P_{ijk} \in H_{ij}$$

where

$$H_{ij} \in S_{im}$$

and S_{im} is the set of all matched households for $i = 1$ to 2. An attempt is made to match all the persons P_{1lr} in household H_{1l} with the corresponding person P_{2ms} in household H_{2m}, where households H_{1l} and H_{2m} have been previously linked for $< = m_{1j}$, and $s < = m_{2j}$.

The Set Matching System matches persons between linked households by using an extension of the Soundex Algorithm to evaluate the first and last names of potentially matched individuals. An age and year of birth sensitivity indicator, similar to an age range, evaluates potentially matched occupants' age similarity. Also considered in the occupant matching process are the sex, race, and marital status of the occupants.

5.3.1 THE SOUNDEX ALGORITHM

The Soundex Algorithm, developed by R.C. Russell, collapses the alphabet into six phonetically unique consonant letters (SSA, 1965a; SSA, 1965b). The purpose of the Soundex Algorithm is to group together names of the same sound, regardless of their actual spelling.

The Set Theory Matching System expands the basic Soundex Algorithm in two areas. First, priority is given to initial soundex consonants, which reduces the influence of suffixes in the name matching process. Secondly, the expanded Soundex Algorithm does not convert the initial letter of the first and last names through the Soundex code substitution process.

5.3.2 THE AGE SENSITIVITY INDICATOR

In addition to the first and last names, individuals are matched on demographic characteristics including age, race, marital status, and sex. Of these four, age is most problematic owing to misreporting, age rounding and, in the case of the Alternative Enumeration, the occurrence of estimated age when reported age was not available. To compensate for the discrepancies in age between the two data sets, an age sensitivity indicator, based on age and year of birth data, is used in lieu of reported age. The sensitivity indicator examines both the age and year of birth fields because of the above-mentioned problems associated with missing or inaccurate data. The age sensitivity indicator considers individuals to be of similar age if they meet the following criteria: within 2 years for persons age 1 to 5; within 5 years for persons age 6 to 25; within 10 years for persons age 26 to 50; or, within 15 years for persons over age 50.

With the exception of age, the household occupant matching algorithm initially follows stringent matching guidelines with regard to name and demographic characteristics. With additional iterations, the matching requirements are relaxed. Matching individuals by name based on the full Soundex consonants and then placing priority on the initial consonants only is an example of this. Relaxing of the individual matching requirements occurs both within a single variable, such as is the case for the name matching procedure, as well as among variables. (A complete explanation of this process is beyond the limits of the paper.)

Similar to the algorithmic reduction of the household linking process, matched household occupants are removed from further matching consideration. Therefore, each pair of matched household occupants (P_{1lr} and P_{2ms}) redefines the subsequent census household universe (H_{1lq}) and Alternative Enumeration household universe (H_{2mq}). That is,

$$H_{1lq} = H_{1l} - P_{1lr}$$

and

$$H_{2ms} = H_{2m} - P_{2ms},$$

where l represents the index of the matched household in S_1, m represents the index of the matched household in S_2, r represents the index of matched person in household H_{1j}, s represents index of the matched person in H_{2j}, and q is the iteration number of the household occupant matching process.

As the iterative household occupant matching process continues, the number of person elements remaining in the sets of linked households (H_{1lq} and H_{2mq}) decreases. Reduction in the number of potential household occupant matches reduces, in turn, the algorithmic requirements necessary for matching subsequent household occupants. This is due to the increasing diversity of remaining unmatched household occupants. Similar to the algorithmic process in household linking, improved match rates and processing speed result from this set reduction method.

5.4 STAGE FOUR: THE OUTSIDE THE HOUSEHOLD INDIVIDUAL MATCHING PROCESS

Household occupants who have not been matched in the previous three stages of the Set Theory Matching System are processed through the outside the household individual matching process. The universe for this pass includes all unmatched individuals regardless of their household link status.

The outside-the-household individual matching process is similar to the more traditional matching systems, except that address elements are not selected as components for matching consideration. Only unmatched individuals remaining in the data sets are considered. Potentially matched individuals are evaluated based on the expanded Soundex algorithm and demographic characteristics. The occupant's household affiliation is disregarded.

6. MATCHING RESULTS

Two methods were used to evaluate the Set Theory Matching System. The initial evaluation compared the output of the purely automated Set Matching System results to a field follow-up study that was conducted by professional anthropologists designed to resolve all discrepancies between the 1990 Census and Alternative Enumeration. The secondary evaluation compared the Set Theory Matching System determinations, augmented by a clerical review procedure, to the field follow-up study. The clerical review procedure corrected matching errors, identified additional matches, and recorded the matching modifications. This evaluation was conducted to determine the accuracy of the complete matching process.

It was a goal of the system development effort to minimize the amount of human intervention necessary in the record linkage process. Therefore, comparing the results of the purely automated Set Theory Matching System with the Total System provides an indication of the effectiveness of the Set Theory Matching System levels.

6.1 FIELD RESEARCH MATCHING RESULTS

A field review, independent of the clerical review, was used to determine the accuracy of the Set Theory Matching System. Professional anthropologists returned to the enumeration site to evaluate the match status of each record. As a result of this review, system error rates--Type I and Type II--were calculated. Type I errors ("false positive") resulted when the system matched two records that were found to be incorrectly matched during clerical review. Type II errors ("false negative") resulted when the system incorrectly classified a record as unmatched when, in fact, a true match was found during field follow-up resolution.

According to clerical review determinations, the Set Theory Matching System yielded an overall accuracy rate (correctly matched plus correctly unmatched records divided by the total number of potentially matched records) of 80.3 percent. The accuracy rate varied from a high of 93.6 percent to a low of 45.8 percent across the 28 sites. Two methods for calculating the overall Type I and Type II error rates are given. The first method is calculated by using the total number of potentially matched records in the denominator. The second method is calculated using only selected records in the denominator, which produces slightly higher Type I and Type II error rates. These calculations yield overall Type I and Type II error rates of 1.4 and 1.8 percent, respectively, and conditional Type I and Type II error rates of 5.3 and 25 percent, respectively. Overall Type I error rates varied from a low of 0.0 percent to a high of 3.4 percent, while conditional Type I errors varied from 0.0 to 10.9 percent. Overall Type II rates ranged from 6.4 to 52.7 percent, while Conditional Type II errors varied from 7 to 68.7 percent.

6.2 TOTAL SYSTEM MATCHING RESULTS

The total system match determinations used the Set Theory Matching System as baseline matching results and augmented the baseline by the clerical review procedure. During clerical review, the match status of each record was manually checked to determine whether it was correctly matched by the computer system. The combined results of using the Set Theory Matching System and Clerical review procedure were compared to the field research results, as was the case in the field research matching results. An overall system accuracy rate, overall Type I and Type II error rates, and conditional Type I and Type II error rates were calculated in the same manner as the field research review. According to the field match results, the total system had an overall accuracy rate of 91.1 percent, Type I error rates of less than 1 percent and 2.8 percent, and Type II error rates of 2.8 and 11.2 percent. As was expected, the results indicate that the clerical review procedure increased the system accuracy and decreased error rates.

6.3 DISCUSSION

The discussion section reviews the contributions that this project has made to record linkage research and the limitations of the Set Theory Matching System. The primary limitations to the system arise from problems with missing and conflicting data contained in the Census and Alternative Enumeration data sets. These system limitations and data problems will be discussed and highlighted with examples from specific enumeration sites.

6.3.1 PROJECT CONTRIBUTIONS TO RECORD LINKAGE RESEARCH

This project has provided a unique contribution to record linkage research for four main reasons: First, a field research reinterview process was conducted in resolving final match determinations. Many record linkage systems, due to funding and time constraints, must rely on statistical models to make final match determinations. Actual field follow-up on 100% of the records provided very stringent criteria to evaluate the Set Theory Matching System.

Secondly, the actual final match database is an invaluable resource which can be used for future record linkage research. The database can be used to prototype, refine, and evaluate future record linkage systems and algorithms. This database may prove to be the an extremely useful tool for continued record linkage development at the U.S. Census Bureau.

Thirdly, this project has evaluated the record linkage process from two standpoints. The initial analysis evaluated the record linkage process from a purely automated standpoint, by comparing Set Theory Matching System results to final field resolution data. The secondary analysis compared the total system, which includes Set Theory Matching System results, augmented by clerical review to the final field resolution data. The dual analysis allows for a comparison of purely automated results and the ability of human beings to enhance the record linkage operation.

Lastly, The Set Theory Matching System used all the data records enumerated in the Alternative Enumeration and Census, regardless of data record completeness. If a person with minimal information was enumerated, the data record qualified for the study. Some record linkage systems set threshold levels for data record quality. Therefore, in some systems, records with missing name field information are disqualified from evaluation procedures.

6.3.2 THE MISSING DATA PROBLEM

Missing data is one of the most significant problems in record linkage research. The Set Theory Matching System has developed a new and effective method to alleviate the missing data problem by using household data structures to simulate human creativity, however the missing data problem will continue permeate record linkage research. When more information is made available, to man or machine, increased match rates and reduced error rates result.

6.3.2.1 THE BUN SITE

In this study, for example, the Alternative Enumeration of the BUN site failed to capture last name information for approximately 70 percent of the enumerated persons, and was missing both last and first name information over 65 percent of the enumerated persons. However, the Set Theory Matching System accurately matched 46 percent of the people at the site with only a 1.5 percent overall type I error rate. The site has an extremely high overall type II error rate of 53 percent. Initially, one may believe that this demonstrates flaws in the record linkage system; however, further analysis shows that the error rate is due the massive amount of unreported Alternative Enumeration information. The field researcher at the site was able to confirm many more matches than were found by the Set Theory Matching System. This is because more information was made available to the researcher through the field follow-up and reinterview process than was contained in either the Alternative Enumeration or Census data sets. The researcher was able to confirm name information that was not available in either data set.

6.3.2.2 THE AMM SITE

The missing information problems is further demonstrated in the AMM site. This site is missing large amounts of data from both the Census and Alternative Enumeration. Over 10 percent of the Alternative Enumeration persons and over 13 percent of the Census persons had missing first and last name information. In addition, 25 percent of the Alternative Enumeration persons and 16 percent of the Census persons had missing age and year of birth information. Clearly, the information contained in the AMM site is limited. Therefore, the Set Theory Matching System and clerical review procedures had difficulty confirming matched pairs. This resulted in relatively low accuracy rates and correspondingly higher overall Type II error rates, however the Overall Type I error rate was held to less than 1 percent.

6.3.2.3 THE VEL SITE

The VEL site is another example of the effects of large amounts of missing name and demographic data. The site has missing name and demographic information on a larger percentage of Census enumerated persons. Over 15 percent of the Census enumerated persons had missing first and last names, over 16 percent had missing year-of-birth and age information, nearly 16 percent had missing sex data, and over 27 percent had missing race determinations. This poor enumeration quality resulted in a comparatively low automated accuracy rate of 54 percent and a correspondingly low total system accuracy rate of 65 percent. The Census data set did not contain enough information to confirm matched pairs of records; therefore Type II error rates in the purely automated system and total systems were much higher than average.

6.3.3 THE CONFLICTING DATA PROBLEM

Conflicting data is another significant problem in record linkage research. The Set Theory Matching System has addressed conflicting data issues in a similar manner to the missing data problem. The system uses records grouped in a household arrangement to provide further insight into solving conflicting data problems. The automated system has attempted to mimic human thought patterns by evaluating record linkage decisions in a household based framework. Conflicting data problems are expected to continue to characterize to record linkage environment, therefore record linkage systems are expected to resolve these problems in an automated manner. The Set Theory Matching System has made significant breakthroughs in record linkage research, however work in solving conflicting data problems is a continuing process.

6.3.3.1 THE ROM, ROD AND DOM SITES

The ROM, ROD and DOM sites are examples of areas where further research is necessary. These sites were characterized by high percentages of conflicting name and race information. The sites had a predominately Hispanic ethnicity and it is believed that enumeration confusion resulted due to the complexity of hispanic naming conversions and the general confusion surrounding race and ethnicity classifications.

Many confirmed matched pairs in the Census and Alternative Enumeration conflicting name information for these sites. This is believed to be particularly prevalent in these predominantly hispanic sites because married hispanic woman have an ethnic tradition of retaining their maiden and husband's surnames. The record linkage process at these sites was further complicated by race classification confusion. For example, in the ROD site the alternative enumeration classified hispanic populations as belonging to the "white" race category, while the census enumerated and mail back returns classified the population as belonging to the "other" race classification. The ROM site was characterized by the reverse race classification situation. Many Alternative Enumerated persons were classified as belonging to the "other" race category, while the Census classified the people as belonging to the "white" race classification.

The name and race data collection confusion resulted in relatively low automated accuracy rates (ranging from 71 to 75 percent) and comparatively higher total system accuracy rates (ranging from 92 to 98 percent). The people performing the clerical review procedure were able to easily resolve the name and race discrepancies, while the automated system lacked algorithmic procedures for confirming matched pairs with conflicting name and race information. The approximately 20 percent increase between the purely automated and total system accuracy rates highlights limitations in the automated system. Algorithm design procedures need to be improved for conflicting name and demographic information, especially for predominately hispanic populations.

7. CONCLUSION

The introduction of household related information, through the use of mathematical set theory, has produced promising initial results in the accuracy of the record linkage process. This paper presented an overview of the Set Theory Matching System and some of the initial research findings. Continued research in evaluating the effectiveness of the Set Theory Matching System is planned for seven main areas.

First, the system will be ported from a DOS/Pascal based platform to a UNIX/C system. This will create a production environment for larger record linkage applications (ie. administrative record linkage), and create a unified record linkage platform for Census Bureau record linkage research.

Secondly, the actual record linkage database can be used to test and refine linkage systems and algorithms. The actual data collected from the field follow-up research can provide an excellent resource for continued system development.

Thirdly, record linkage accuracy rates can be correlated with missing data elements to evaluate the effect of missing data on Set Theory Matching System results. This will lead to continued algorithm refinements.

Fourth, an audit trail that would identify the precise process location of a system determined match could be used to further refine the system. The audit trail would provide information for further system refinement by correlating error rates with specific process algorithms. In addition, the audit trail could be used to evaluate the differences between missing and conflicting data at various stages of the matching process.

Fifth, a detailed comparative analysis of the Set Theory Matching System with traditional matching systems would provide additional insight into the record matching process. Merging the methodologies of several record matching systems could augment the strengths of each independent system.

Sixth, an evaluation could be conducted which isolates the advantages of household record linkage elements. The Set Theory Matching System, once ported to the UNIX/C platform, could be adapted to run without the household level component. This would allow a detailed comparison of the advantages of household incorporated set theory with more traditional record linkage systems and provide further confirmation of the advantages of household based information into the record linkage process.

Lastly, an ethnographic evaluation of the Set Theory Matching System based on demographic and household characteristics may offer avenues for continued system refinements. For example, the heavy concentration of Chinese immigrants at the SUN site may warrant different matching criteria based on ethnicity.

Notes

1. Type I errors are those in which a record was incorrectly matched by the Set Theory Matching System.

2. Type II error are those in which the Set Theory Matching System incorrectly classified a record as unmatched when, in fact, a true match did exist.

REFERENCES

Fellegi, I. and Sunter, A. (1969), "A Theory for Record Linkage," Journal of the American Statistical Association, 40.

Horowitz, E. and Sahni, S. (1987), Fundamentals of Data Structures in Pascal, Rockville, Maryland: Computer Science Press.

Kernighan, B and Ritchie, D. (1988), The C Programming Language, Englewood Cliffs, New Jersey: Prentice Hall.

Social Security Administration (1965a), Central Office Bulletin, Vol. 2, "Soundex Code History.", February 12, 1965.

Social Security Administration (1965b), Central Office Bulletin, Vol. 2, "DAO's Coded File.", February 19, 1965.

Statistics Canada/Systems Development Division (1982), "Record Linkage Software."

Statistics Canada/EDP Planning and Support Division (1984), "Record Linkage Software."

Winkler, W. (1985), "Preprocessing of Lists and String Comparisons" in Record Linkage Techniques-1985, edited by W. Alvey and B. Kilss, U.S. Internal Revenue Service Publication 1299 (2-86), 181-187.

Winkler, W. (1987a) "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," Technical Report.

Winkler, W. (1987b) "Computational Aspects of Applying of the Fellegi-Sunter Model of Record Linkage to Lists of Businesses" Statistical Uses of the Administrative Data Proceedings, November 1987.

Wirth, N. (1976), Algorithms + Data Structures = Programs, Englewood Cliffs, New Jersey: Prentice Hall.