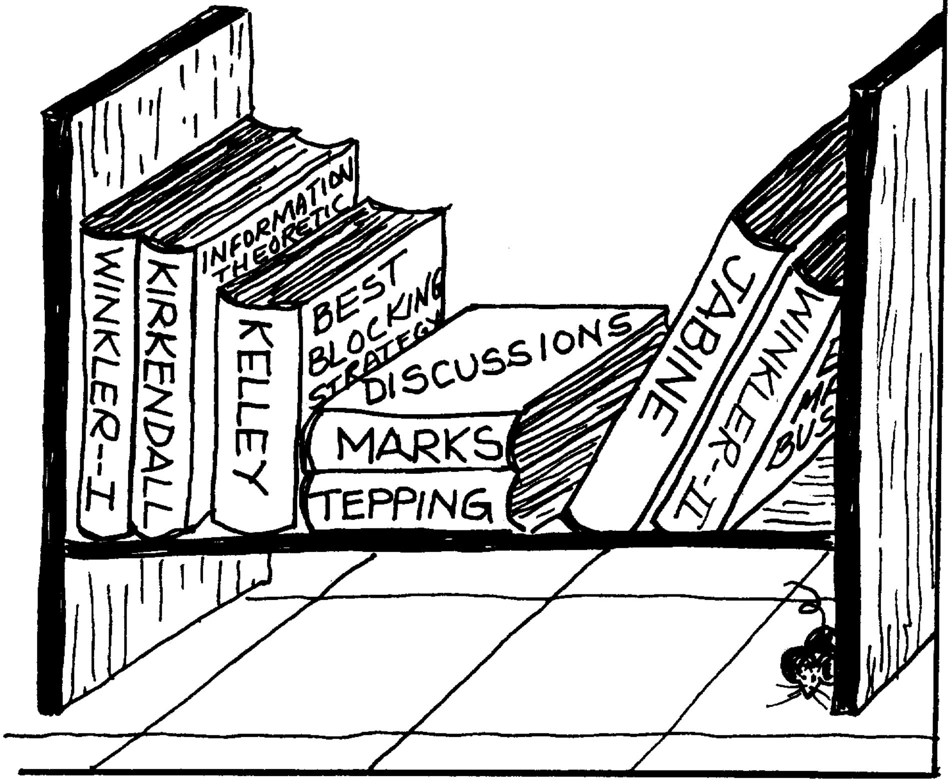
Section III: Current Theory and Practice



PREPROCESSING OF LISTS AND STRING COMPARISON

William E. Winkler, Energy Information Administration

1. INTRODUCTION

By combining data on entities from different sources, researchers are often able to perform analyses that would not be possible if they were to use data from individual sources separately.

When a unique common identifier (such as a verified Social Security Number) is available on individual sources of data, matching files merely involves using the unique identifier as the sort key and then directly matching records from the two files.

When a unique common identifier is not available, it is necessary to use other identifying information. Characteristic identifying information might consist of surname, street address, or ZIP code in matching files that contain name and address information. Use of such information involves several practical problems.

First, if the precise locations of identifiers (such as first name and surname) are not consistent from record to record, computer matching using the identifiers cannot be performed. Second, some identifiers may be miscoded or missing on some records. Third, such identifiers, or even combinations of them, are not unique for individuals or businesses.

This paper presents examples of some of the solutions for problems arising in preparing name and address information for use in matching files.

Most of the work described has taken place at the U.S. Bureau of the Census, the Statistical Reporting Service in the U.S. Department of Agriculture, the Energy Information Administration, and Statistics Canada. The problems, examples, and resultant methodologies should be representative of problems that arise in general.

2. BACKGROUND

2.1. Why Preprocessing is Needed

Match/merge strategies generally perform better (i.e., have lower rates of erroneous matches and nonmatches) when address lists have been preprocessed to produce more consistent formats and spellings and to delineate records representing different types of entities (such as records associated with individuals/ sole proprietorships, partnerships, and businesses).

2.2. Definitions

As the terminology of matching is not always consistent from reference to reference, we present definitions.

A match is a pair of records that represent the same unit and a nonmatch is a pair of records that do not. Blocking is a procedure for subdividing files into a set of mutually exclusive subsets under the assumption that no matches occur across blocks. Each mutually exclusive subset consists of records agreeing on the blocking characteristics.

A positive link is a pair of records that is designated as a match. A positive nonlink is a pair of records that is designated as a nonmatch. A possible link is a pair of records that is not designated as a positive link or nonlink. Additional steps, such as manual review or collection of additional information, are needed to designate it as a positive link or nonlink.

A Type I Error is the designation of a pair of records as a positive nonlink when it is a match. Type I Errors have been referred to as erroneous or false nonmatches (U.S. Department of Commerce, 1980). A Type II Error is the designation of a pair of records as a positive link when it is a nonmatch. Type II Errors have been referred to as erroneous or false matches.

2.3. Nature of the Problem

The specific types of match/merge procedures adopted depend on the identifiability and consistency of corresponding information in the address lists to be merged. For instance, if an address list were in free format, then merging would have to be done manually because computer software could not use corresponding information such as NAME or ZIP for blocking pairs of records.

Even if fields such as NAME, ADDRESS, CITY, STATE, and ZIP are identified (possibly using manual techniques), it may not be possible to block records accurately if words in corresponding fields do not contain consistent spellings. For instance, the STATE field and words such as 'COMPANY,' 'CORPORATION,' 'P O BOX,' and 'STREET' should be spelled or abbreviated in a consistent manner.

If subfields such as FIRST NAME, MIDDLE INITIAL(S), SURNAME, STREET NUMBER, STREET NAME, PO BOX NUMBER, ROUTE NUMBER, and SUITE NUMBER are identified and placed in fixed locations, then they can be used for delineating true and false matches. If FIRST NAME and SURNAME subfields are in inconsistent order within the NAME fields of two lists, then it will not be possible to block records accurately using the NAME field.

2.4. Match/Merge Stages

As the need for specific types of preprocessing is closely connected to different match/ merge strategies, these strategies and their relationship to specific data needs will be summarized.

Matching records within or across lists consists of two stages. In the blocking stage, pairs of records are blocked into sets of pairs using a few common characteristics with substantial discriminating power. Some such characteristics are the SOUNDEX abbreviation of SURNAME (see e.g. Bourne and Ford (1961)) or ZIP code. Records for which such common characteristics do not agree are assumed to represent different entities.

In the discrimination stage, blocked pairs are categorized as positive links, positive nonlinks, or potential links using all available discriminating characteristics within blocked pairs of records.

At both stages preprocessing can play an important role. For instance, if records of individuals are blocked using the SOUNDEX abbreviation of the surname, the location of surname needs to be identified and the spelling of surnames needs to be moderately accurate. If records of establishments or businesses are blocked using ZIP code, then ZIP codes need to be accurate.

If the first name, first four characters of the street address, and state abbreviation are used for designating links and nonlinks within a set of blocked pairs, then those fields and subfields need to be located and accurate.

2.5. Topics Addressed in Paper

The remainder of this paper presents examples of the kinds of name and address lists that are encountered and the types of preprocessing that are performed. The third section presents examples illustrating problems with names and addresses in lists that are normally available for updating. The fourth section presents a summary of the various types of preprocessing software and procedures to identify different types of entities, clean up fields and subfields, and identify subfields of the NAME and STREET ADDRESS fields.

The fifth section describes methods for comparing strings that are used to overcome some spelling variations and to create sort keys. The final section poses some problems for further research.

3. EXAMPLES OF PROBLEMS IN NAME AND ADDRESS LISTS

In addition to the problem of locating sources of lists for use in updating, there are problems associated with lists that can make them difficult to use. Problems can include transferral of hardcopy lists to computer files, identification of fields and subfields, and different name and/or address representation of similar entities or similar representation of different entities.

This section provides examples of the problems that affect a list's suitability for use as an update source.

3.1. Keypunch Error in Consistently Formatted Subfields

Addresses in a source list might contain a significant number of typographical errors -- which do not seriously affect manual processing -- while the computerized mailing list does not. The following two pairs of names and addresses representing two entities, from source lists and mailing lists being updated, respectively, illustrate the problem.

(a)	J K Smoth	114 E Main Stret
	J K Smith	114 Main St
(b)	Southside Feul	898 Northwst Hghwy
	Soth Side Fuel	8895 Northwest Hwy

3.2. Unidentified Fields

Address records in which the five fields NAME, STREET, CITY, STATE, and ZIP occur in free format generally cannot be placed in consistent formats using straightforward computer code. They must be reformatted manually. Free format records often exist as address labels in which the five fields occur in no fixed format.

The following examples illustrate the problem of free formats:

(a) A A Fuel Oil
c/o Marvel Distribution Co
PO Box 519
Laramie, Wyoming 66519

(b) Smith Distributing
5632 Westheimer
Suite 43
Houston TX 77514

(c) ABC Oil, PO Box 54
Grand Rapids
Michigan 49506

In example (a) the name occurs on the second line whereas in examples (b) and (c) it occurs on the first. The STREET/PO BOX field appears on the third, second, and first lines of examples (a), (b), and (c), respectively. The CITY field appears in the second to last line in example (c) but on the last line in examples (a) and (b).

3.3. Inconsistently Formatted Subfields

If formatting conventions within subfields of the name and address field vary substantially, merging procedures may not perform as well as in the situation in which corresponding subfields can be readily identified using computer software. For instance, one or more lists might contain records with names and addresses in the following forms:

(a)	J K Smith Co	113 Main
	Smith J K Co	113 E Main St
	Smith Jonathon K Co	PO Box 16
(b)	A A Fuel Co	PO Box 105
	AA Fuel Distribution Inc	Drawer 105
(c)	R Smith Fuel Co	1171 Northwest Highway
	Robert Smith	Highway 65 West
	Smith Co	Route 1

In the first two lines of example (a), both SURNAME and STREET NAME are not obvious matches using a straightforward computer comparison and the billing address in the third entry makes it difficult to determine if the three entries represent the same company.

In example (b), the COMPANY NAME subfields cannot be easily identified and the ADDRESS fields may be difficult to compare. In the example (c), SURNAMES may not be identified and the equating of street addresses of the first two entries requires specific geographic information. Without additional information, it is difficult to determine whether the third entry represents the same company as that given by the first two entries.

3.4. Name and Address Representation

3.4.1. Same Entity, Different Name and Address

Entities in some potential update sources are represented in substantially different forms than the entities are represented in the main mailing list. When this happens, it is difficult to determine those records representing entities that are out-of-scope or duplicates to records in the main mailing list.

For instance, a list of individuals licensed by a state to sell petroleum products might be considered as an update source for a list of businesses selling petroleum products in the state. The reason that the list of owners might be considered is that sending a form to either the owner of a small fuel oil dealership or the appropriate corporate billing address (which might exist in the main mailing list) could yield correct sales information.

Combining such a list of owners with a list of businesses can yield difficulties. Without a suitable additional data source, it may be impossible to identify records representing the same entity that take the following form:

J K Smith 116 Main St
Anytown 66591
A A Fuel PO Box 68
Othertown 66442

3.4.2. Same or Different Entity, Similar Name, Different Address

If the purpose of a mailing list is to provide one address record for each corporate entity, then additional difficulties can arise. Businesses often maintain substantially different mailing addresses, sometimes even requiring survey forms to be sent to locations in different states. For instance, addresses could take the following form:

ABC Fuel Co 116 Main St
Anytown CA 96591
ABC Fuel Oil PO Box 534
Othertown NY 10091
J K Smith ABC Co PO Box 68
Sometown KS 66442

The first two records could represent the same corporate entity, independent but affiliated companies, or unaffiliated companies. The third address could represent a subsidiary of one of the companies represented by the first two records, a subsidiary of an unidentified company, or an affiliated but independent distributor of products for some ABC Co.

3.4.3. Different Entity, Identical Address and/or Phone

With some lists, different entities may be represented as follows:

(a) Pargas of Illinois PO BOX 661
NY 10015 202/664-2139
Pargas of Ohio PO BOX 661
NY 10015 202/664-2139

(b) ABC Distributing 1345 Westheimer
TX 71053 703/789-5439
Lone Star Oil 1345 Westheimer
TX 71053 703/789-5439

Example (a) illustrates a situation in which a parent company reports separately for two subsidiaries. Example (b) could represent a situation in which an accountant reports for two different companies. The address and phone number could be the accountant's.

Example (b) could also represent different companies which are both located in the same office building or two different companies, one of which has gone out of business. If companies are matched using TELEPHONE, manual followup may be required to determine whether one has gone out of business or is an affiliate of the other.

4. PREPROCESSING METHODS

Methods of preprocessing, using manual procedures or software, have been developed to (1) delineate corresponding classes of records such as those associated with corporations, partnerships, or individuals within a list of businesses; (2) identify corresponding subfields such as HOUSE NUMBER, STREET NAME, and PO BOX; (3) make consistent the spelling of words such as 'STREET,' 'CORPORATION,' and 'ROUTE;' and (4) clean up ZIP codes.

4.1. Identification of Individuals, Partnerships, and Corporations

As records associated with individuals/sole proprietorships, partnerships, and corporations within a list of businesses have different characteristics, they are sometimes distinguished and processed separately. The U.S. Department of Agriculture/Statistical Reporting Service (USDA/SRS, 1979) and the U.S. Department of Commerce (1981) have developed software and/or procedures for identifying individuals, partnerships, and corporations in lists of farms.

It appears that partnerships are identified as those records having '&' in the NAME field. Corporations are those records having words such as 'CORP,' 'CO,' 'INC,' 'FARMS,' and 'DAIRY' in the NAME field. Individuals are those records not classified as partnerships or corporations.

Records associated with partnerships are more difficult to process (may require more manual followup) because partnerships can be erroneously matched more times than records associated with individuals and because partnership records can take the following inconsistent forms:

Smith John A & Mary B
Smith John & Jones Lee
Smith John A, Smith Mary B, & Lee Jones
Smith Mary B & Jones Lee
Smith Mary B & Smith John A

The first entry contains only one SURNAME entry while others contain one SURNAME for each partner. The third entry represents a partnership of three individuals while the others represent only two. Due to ordering differences in entries two through four, it is difficult to determine if Jones or Lee is the individual's surname.

4.2. Formatting and CLeanup of the Name Field Subfields

Cleanup of the name field consists of replacing common words such as 'COMPANY,' 'INCORPORATED,' 'LIMITED,' 'FARMS,' 'BROTHERS,' 'SALES,' and 'DISTRIBUTOR' with standard spellings or abbreviations and replacing common variations of first names such as 'ROBERT,' 'BOB,' 'ROB,' 'ROBT' with standard spellings or abbreviations.

The standardization is typically done using lookup tables that contain previously identified spelling variations. Such lookup tables are easily updated when new spelling variations are encountered. Lookup tables are in use at USDA/SRS (1979), the U.S. Department of Commerce (1978b, 1981), the Energy Information Administration (EIA) (Winkler, 1984), and Statistics Canada (1982).

Formatting of name fields associated with individuals involves manually identifying the subfields FIRST NAME, MIDDLE INITIAL(S), and SURNAME and either placing them in fixed locations (USDA/SRS, 1979) or in fixed order (U.S. Dept. of Commerce, 1981). If NAME subfields are in fixed order, then software can be used to identify individual subfields.

4.3. Formatting and Cleanup of the Street/Mailing Address Field

Cleanup of the street/mailing address involves replacing such commonly occurring words as 'STREET,' 'PO BOX,' 'RURAL ROUTE,' 'DRAWER,' 'AVENUE,' and 'HIGHWAY' with standard spellings or abbreviations. Such standardization typically involves lookup tables that are easily updated as new spelling variations are encountered.

Various spellings of large cities in the CITY field can also be standardized using lookup tables. Such standardization may only be partially effective because of the large differences in spelling and abbreviations used for core cities and suburbs in large metropolitan areas.

Formatting can also involve placing subfields such as STREET NAME, STREET NUMBER, PO BOX NUMBER, RURAL ROUTE in fixed locations (USDA/SRS, 1979; U.S. Dept. of Commerce, 1978b; Statistics Canada, 1982).

ZIPSTAN software (U.S. Dept. of Commerce, 1978b) has been developed to identify pertinent subfields of the STREET field in files of individuals. The following examples show representative EIA records before and after ZIPSTAN processing:

Figure 1. -- Before ZIPSTAN

1. EXCH ST
2. HWY 17 S
3. 1435 BANK OF THE
4. 2837 ROE BLVD
5. MAIN & ELM STS
6. CORNER OF MAIN & ELM
7. 100 N COURT SQ
8. 100 COURT SQ SUITE 167
9. 2589 WILLIAMS DR APT 6
10. 15 RAILROAD AVE
11. 2ND AVE HWY 10 W
12. MAIN ST
13. 184 N DU PONT PKWY
14. 1230 16TH ST
15. BOX 480

Figure 2. — After ZIPSTAN

No.	House No.	Prefixes 1	Prefixes 2	Street Name	Suffixes 1	Suffixes 2	Unit
1.				EXCH	ST		
2.		HW		17TH	S		
3.	1435			BANK OF THE			
4.	2837			ROE	BL		
5.				MAIN ELM STS			
6.				CORNER OF MAIN ELM			
7.	100	N		COURT	SQ		
8.	100	CT	SQ	*** NO NAME ***			RM 167
9.	2589			WILLIAMS	DR		AP 6
10.	15			RAILROAD	AV		
11.				2ND	AV	HW	10
12.				MAIN	ST		
13.	184	N		DU PONT	PW		
14.	1230			16TH	ST		
15.	480			*PO BOX*			

ZIPSTAN is able to identify accurately subfields in 13 of 15 cases. The two exceptions are cases 2 and 8. In case 2, 'HWY' is moved to a prefix position and '17' is placed in the STREET NAME position. In case 8, 'COURT,' the STREET NAME, is placed in a prefix location.

Although ZIPSTAN accurately identifies the subfields associated with intersections (cases 5, 6, and 11), such identification may not allow accurate delineation of duplicates in comparisons of various lists. Some lists may contain STREET ADDRESS in the following forms, none of which is readily comparable with the forms in examples 5, 6, and 11.

5. 34 Main St
5. Elm and Main Streets
11. Hwy 10 W
11. 7456 Richmond Hwy

5. METHODS OF STRING COMPARISON

If comparable strings have been identified (see sections 3.4, 4.2, and 4.3), then it is useful to compute a distance between them in blocked pairs of records. If properly devised, string comparators can overcome minor spelling errors.

5.1. Abbreviation Methods

Abbreviation methods (see e.g., Bourne and Ford, 1961) are intended to maintain some information needed for identifying a record while alleviating problems due to spelling variations. As an example, the SOUNDEX abbreviation method will be described and illustrated.

The SOUNDEX abbreviation of an alphabetic word consists of four characters. The first SOUNDEX character agrees with the first character in the word. All nonleading vowels and the letters H, W and Y are deleted. Similar sounding consonants are mapped into integer codes as follows:

B, F, P, V -> 1,
C, G, J, K, Q, S, X, Z -> 2,
D, T -> 3,
L -> 4,
M, N -> 5, and
R -> 6.

Repeating integer codes are deleted and SOUNDEX abbreviations of less than four characters are zero filled on the right.

Comparison of SOUNDEX abbreviations of words induces a metric in which agreeing SOUNDEX abbreviations are assigned distance 0 and disagreeing 1.

5.2. General String Comparators

As common abbreviation methods (section 5.1) are not able to deal with typical coding errors, more exotic methods for string comparison have been introduced.

An early comparator is the Damerau-Levenstein (D-L) metric (see e.g., Hall and Dowling, 1980, pp. 388-390). The basic idea of the metric is as follows. Any string can be transformed into another string through a sequence of changes via substitutions, deletions, insertions, and possibly reversals. The smallest number of such operations required to change one string into another is the measure of the difference between them.

The minimum value that the D-L metric can assume is 0 (character-by-character agreement) and the maximum is the maximum number of letters in the two words being compared. For instance, the D-L distance between 'ABCDEFG' and 'WXYZ' is 7.

Using the Damerau-Levenstein metric or various straightforward extensions of it (see e.g., Hall and Dowling, 1980) is difficult because: (1) the dynamic programming necessary for computing the metric is cumbersome and (2) neighborhoods of given strings contain too many unrelated strings (i.e., the metric does not have good distinguishing power, see section 5.3).

5.3. Jaro's String Comparator

Jaro (see e.g., U.S. Dept. of Commerce, 1978a, pp. 83-108) introduced a string comparator that is more straightforward to implement than the Damerau-Levenstein metric and more closely relates to the type of decisions a human being would make in comparing strings.

The string comparator is a weighting function for pairs of strings denoted as reference file strings and data file strings. It is defined as follows (U.S. Dept. of Commerce, 1978a, p. 108):

$$W = wgt\_cd*c/d + wgt\_rd*c/r + wgt\_tr*(c-tr)/c$$

where

wgt_cd = weight associated with characters in the data file string but not in the reference file string;

wgt_rd = weight associated with characters in the reference file string but not in the data file string;

wgt_tr = weight associated with transpositions;

d = length of the data file string;

r = length of the reference file string;

tr = number of transpositions of characters; and

c = number of characters in common in the two strings.

Two characters are considered in common only if they are no further apart than (m/2 - 1) where m = max(d,r). Characters in common from two strings are said to be assigned. Other characters from the two strings are unassigned. Each string has the same number of assigned characters because each assigned character represents a match.

The number of transpositions are computed as follows: The first assigned character on one string is compared to the first assigned character on the other string. If the characters are not the same, half of a transposition has occurred. Then the second assigned character on one string is compared to the second assigned character on the other string, etc. The number of mismatched characters is divided by two to yield the number of transpositions.

If two strings agree on a character-by-character basis, then the Jaro weight, W, is set equal to wgt_cd+wgt_rd+wgt_tr, which is the maximum value that W can assume. The minimum value that the Jaro weight, W, can assume is 0, which occurs when the two strings being compared have no characters in common (subject to the above definition of common).

5.4. Manual Comparison

The purpose of different string comparators is to assign a value to the quality of comparison in a manner that mimics how a human being might make a decision. Because of this, it is useful to describe how manual review decisions can be quantified. In section 5.5, the manual review decisions will be compared to results obtained using the string comparators of sections 5.1-5.3.

Quantification of manual review decisions can be performed as follows:

1. have a number of individuals compare pairs of corresponding substrings such as SURNAMEs;
2. score comparisons using the scale: 1-no match, 2-likely false match, 3-possible true match, 4-likely true match, and 5-true match; and
3. average results of the comparisons over individuals and compute the corresponding coefficients of variation.

5.5. Comparison of String Comparators

Table 1 provides a comparison of the measures of agreement using the SOUNDEX abbreviation, the Damerau-Levenstein metric, Jaro's string comparator, and a weight based on manual review. To make the values in the table easier to compare, all measures were transformed to a scale from 0 to 1. A value of 0 represents nonmatch and a value of 1 represents match.

The transformations are performed as follows:

1. SOUNDEX=1-SOUNDEX;
2. D_L =(5-D_L)/5;
3. JARO =JARO/900; and
4. MAN =(MAN-1)/4.

In equations 1-4 the measures on the right-hand side (as defined in sections 5.1-5.4) are replaced by the scaled measures. As the basic Damerau-Levenstein metric D-L (section 5.2) on the right-hand side of equation 2 varies from 0 (total agreement) to 5 (substantial disagreement) for the examples in Table 1, the scaled

D-L metric is transformed into a weight in which 0 and 1 represent nonmatch and match, respectively.

In computing the Jaro weight, JARO, the weights wgt_cd, wgt_rd, and wgt_tr (section 5.3) are each given the values 300 which are the same as the default values given in the Census software (U.S. Dept. of Commerce, 1978a, p. 88). As the basic JARO weight on the right hand side of equation 3 varies between 0 and 900, dividing by 900 changes the scale from 0 to 1.

In Table 1, with the exception of example (h) (completely different words), all examples represent similar character strings that disagree because of minor transcription/keypunch errors. Each pair of surnames is taken from EIA files. With the exception of example (h), the surnames represent the same entity.

Overall, we can see that the SOUNDEX weight is high for only 5 of 9 matching surname pairs; D-L weights are generally moderately high to high for 8 of 9; Jaro weights are consistently high; and the manually estimated weights vary significantly with no apparent consistency. It is important to note that, with the exception of example (h), all weights should be consistently high.

In comparing the D-L metric and the Jaro weight, we see that the Jaro weight gives additional weight to longer, but similar, strings. For instance, with short strings in which one character disagrees (examples (f) and (i)), the D-L and Jaro weights are about the same. With longer strings in which one character disagrees (examples (d) and (e)), the Jaro weight is higher than the D-L weight.

For example (g), it is interesting to note that the manually estimated weight of 0.88 is lower than the weight of 1.0 provided by each of the other string comparators. Human beings are able to make use of the auxiliary information that "Smith" is a commonly-occurring word and downweight their judgements accordingly. Such downweighting is inherent in the application of the Fellegi-Sunter model which utilizes frequency of occurrence of character strings (see e.g., Rogot, Schwartz, O'Conor, and Olsen, 1983, p. 324).

6. NEEDED FUTURE WORK

Although it is intuitive that preprocessing can both identify information that should correspond and make such information more consistent, few, if any, studies have been set up to determine its effectiveness. We do not know how much different types of preprocessing reduce matching error rates, nor do we know the extent to which they lower amounts of manual processing.

Effective evaluation may require the creation of data bases with all matches identified and suitably connected to entities used for mailing purposes. Fellegi and Sunter (1969) indicate that error rates obtained using samples are subject to substantial variability unless the samples are very large. Winkler (1984) provides examples of rates of erroneous nonmatches based on samples of size 1,800 for which the estimated sampling error exceeds the estimated error rate.

A key issue that needs to be addressed is whether the results obtained by empirical evaluation of methodologies on one data set are likely to be relevant to a different data set. Specific research problems follow.

6.1. Effects of Spelling Standardization

How much does standardization of the spelling of words such as 'COMPANY,' 'CORPORATION,' 'PO BOX,' 'STREET,' and 'EAST' reduce the error rates associated with a given matching strategy? What errors can certain types of standardization induce?

Some matching strategies consist of blocking files of individuals using the SOUNDEX or New York State Intelligence and Identification (for NYSIIS, see Lynch and Arends, 1977) abbreviations of surnames. When compared with blocking using surname, how much does blocking using abbreviated surnames reduce the rate of erroneous nonmatches and can such abbreviations provide information useful for delineating matches and nonmatches within the set of blocked pairs?

Some matching strategies consist of blocking files of businesses using the ZIP code and first few characters of the NAME field. How much effort is involved in cleaning up ZIP codes and how much do the cleaner ZIP codes reduce rates of erroneous nonmatches? Should the ZIP codes in a given metropolitan area all be mapped into one sort key used for blocking records?

How much can the delineation of true and false matches be improved if the spelling and formatting of the CITY field are made more consistent? What are the best strategies for correcting inconsistencies in the CITY field?

6.2. Effect of Formatting of Subfields

How much does the identification of SURNAME, FIRST NAME, HOUSE NUMBER, STREET NAME, and PO BOX help reduce error rates? What subfields provide the greatest reduction? Are the subfields providing the greatest reduction different in files of businesses than in files of individuals?

6.3. Abbreviation Methods Used in Blocking

What are the best methods for blocking files of individuals? Blocking on surnames abbreviated using methods such as SOUNDEX and NYSIIS will usually designate as nonmatches those matches containing errors due to miskeying, insertions, deletions, and transpositions.

In comparing methods of abbreviation and blocking, we need to consider rates of erroneous nonmatches, total number of pairs in all blocks, and computing requirements if some blocks are large. Given these evaluation criteria, are there methods of abbreviation and blocking that would perform better than SOUNDEX or NYSIIS?

6.4. Effect of String Comparison

How much does the string comparator of Jaro (section 5.3) that is used for computing agreement weights for corresponding subfields such as SURNAME, FIRST NAME, and STREET NUMBER (U.S. Dept. of Commerce, 1978a) help reduce rates of erroneous matches? Are there better algorithms for string comparison? What measures should be used in comparing the effectiveness of two string comparators?

ACKNOWLEDGMENTS

The author wishes to thank Yvonne M. Bishop, Thomas Jabine, and Nancy J. Kirkendall for a number of helpful comments.

REFERENCES

Bourne, C. P., and Ford, D. J. (1961), "A Study of Methods for Systematically Abbreviating English Words and Names," J. ACM 8, 538-552.

Damerau, F. J. (1964), "A Technique for Computer Detection and Correction of Spelling Errors," Communications of the ACM. 7, 171-176.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," JASA 40, 1183-1210.

Hall, P. A. V. and Dowling, G. R. (1980), "Approximate String Matching," Computing Surveys 12, 381-402.

Lynch, B. T. and Arends, W. L. (1977), "Selection of a Surname Coding Procedure for the SRS Record Linkage System," U.S. Department of Agriculture, Statistical Reporting Service.

Morgan, H. L. (1970), "Spelling Correction in Systems Programs," Communications of the ACM, 13, 90-94.

Newcombe, H.B. and Kennedy, J.M. (1962), "Record Linkage," Communications of the ACM, 5, 563-566.

Rogot, E., Schwartz, S., O'Conor, K., and Olsen, C. (1983), "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index." ASA 1983 Proceedings of the Section on Survey Research Methods, 319-324.

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

U. S. Department of Agriculture/ Statistical Reporting Service (1979), "List Frame Development: Procedures and Software."

U. S. Department of Commerce, Bureau of the Census/Agriculture Division (1981), "Record Linkage for Development of the 1978 Census of Agriculture Mailing List."

U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978a), "UNIMATCH: A Record Linkage System."

U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978b), "ZIPSTAN: Generalized Address Standardizer."

U. S. Department of Commerce, Office of Federal Statistical Policy and Standards (1980), "Statistical Policy Working Paper 5: Report on Exact and Statistical Matching Techniques."

Winkler, W. E. (1984), "Issues in Developing Frame Matching Procedures: Exact Matching Using Elementary Techniques." Presented to the ASA Committee on Energy Statistics in April 1984. A summary appeared in Statistics of Income and Related Administrative Record Research: 1984 U.S. Department of the Treasury, Internal Revenue Service, Statistics of Income Division, 171-176. The summary also appeared in the ASA 1984 Proceedings of the Section on Survey Research Methods, 327-332.

Table 1: Comparison of String Comparator Metrics Using Surnames that are Generally Similar

	Surnames	Maximum string length	SOUNDEX	D-L	Jaro	Manual	CV 1/
(a)	Tranisano Traivsano	9	0.00	0.60	0.93	0.35	40.3
(b)	Alexander Aleander	9	0.00	0.80	0.96	0.63	15.1
(c)	Nuzinsky Newzinski	9	1.00	0.40	0.81	0.42	39.2
(d)	Smthfield Smithfeld	9	1.00	0.60	0.93	0.63	20.2
(e)	Bachman Bahcman	8	1.00	0.80	0.96	0.63	30.9
(f)	Dixon Nixon	5	0.00	0.80	0.87	0.13	35.1
(g)	Smith Smith	5	1.00	1.00	1.00	0.88	24.0
(h)	Smith Jones	5	0.00	0.00	0.00	0.00	0.0
(i)	Ouid Ovid	4	0.00	0.80	0.83	0.55	13.2
(j	Boc Boco	4	1.00	0.80	0.92	0.32	29.3
Number of values above 0.5		NA	5	8	9	5	NA

1/ Coefficient of variation associated with estimate based on manual review by nine individuals.

WEIGHTS IN COMPUTER MATCHING: APPLICATIONS AND AN INFORMATION THEORETIC POINT OF VIEW

Nancy J. Kirkendall, Energy Information Administration

This paper summarizes the historical development of computerized match/merge procedures and describes the test statistic used to classify record pairs as a match or nonmatch in terms of its information theoretic interpretation. Current match/merge software procedures are compared and contrasted based on their differing approaches to estimation.

INTRODUCTION

The match/merge procedures discussed in this paper are those which are intended to perform exact matching. Exact matching has been defined (U.S. Department of Commerce, 1980) as the linkage of records from two or more files containing units from the same population. The intention of exact matching is to link data for the same unit (e.g., person) from different files. If units which do not represent the same individual are linked, the result is a false match or type 2 error. If units which do represent the same unit are not linked, the result is a missed match, or type 1 error.

There are many different purposes in exact matching. Examples range from obtaining more data elements for an individual by merging information from different surveys, to creating a more comprehensive name and address list by merging the names and addresses from many sources. In the first case, it is important to make sure that matching is done accurately so that the merged data constitute a multivariate observation from a single individual (see Kelley, 1983). In the second case, the merging is intended to ensure as complete a list as possible while eliminating duplication.

The most significant paper on the theory and practice of matching is by Fellegi and Sunter (1969). Their paper documents the derivation of a test statistic and a critical region for deciding whether or not a pair of records is a match. In addition, it discusses some of the assumptions necessary for practical application and describes approaches for estimating the probabilities which are used to calculate the test statistic. Most of the probabilistic match/merge procedures in use today are based on an application of the techniques described in the Fellegi-Sunter paper.

Although the Fellegi-Sunter paper was the first publication of the theoretical background for match/merge procedures, many of the ideas and techniques embodied in the methodology had been used since the late 1950's by Howard Newcombe et al. Newcombe's papers from that time period describe the use of the test statistic for which the derivation was later presented by Fellegi and Sunter. (See Newcombe et al., 1959 and Newcombe and Kennedy, 1962.)

THEORETICAL BACKGROUND

Assume that two files, A and B, are to be merged. Each file contains at least one record for each unit (person or establishment) in the file. Each record contains a set of attributes for that unit. These attributes may include: numerical identifiers with very good identifying characteristics such as the social security number; standard identifiers such as name and address; characteristic information such as sex or date of birth; or any other data which might be available on survey files or administrative record files.

In the matching process, each record in file A can be compared to each record in file B. The comparison of any such pair of records can be viewed as a set of outcomes, each of which is the result of comparing a specific attribute from the record in file A with the same attribute in the record from file B. Outcomes may be defined as specifically as desired. For example, one might define an outcome of a comparison to be simply that the attributes agree or that they disagree. Or, one might define the agreement outcome more specifically, based on the possible values that attribute can take. For example, one outcome might be that the surnames agree and equal "Smith," while another might be that the surnames agree and equal "Zebra," etc.

"Comparison of attributes" is usually interpreted to mean that the same attribute is recorded on each record and that they can be compared directly. However, it is possible to "compare" different attributes which are known to be correlated or to use information from only one record in conjunction with general information from the other file. An example is given in Smith, Newcombe, and Dewar (1983). In their application, records from a file of patients diagnosed as having cancer are linked with records in a death file. The variable "cause of death" in the death file is used in conjunction with general statistics concerning the cause of death among cancer patients and the cause of death among the general population to provide a different sort of "comparison of attributes."

In the above, it was implied that every record from file A is compared to every record from file B. In practice, with large files this would require an extremely large number of comparisons, the vast majority of which would not be matches. To make the size of the problem more manageable, files are generally "blocked" using one or more of the available attributes, and record pairs are assumed to be a possible match and subject to the detailed attribute comparison only if they agree on the blocking attribute. In using a blocking procedure, there is necessarily a higher rate of unmatched duplicates (type 2 error) because records which do represent the same unit, but disagree on the blocking attribute, are automatically rejected as possible matches. However, the gains in the form of reduced processing are significant. See Kelley (1985) for a probabilistic approach to selecting blocking strategies.

THE PROBLEM

Probabilistic test procedures are based on evaluating record pairs one at a time and subjecting each pair to a decision as to its match status. The procedure does not consider the expected number of matches or nonmatches in a merging of two files, and does not make use of the result of the classification of any previous record pairs.

In this section the test statistic and the critical region are described based on an information theoretic argument. Details of the derivation are presented in the Appendix. The resulting test statistic and critical region are exactly the same as those derived by Fellegi and Sunter. One advantage of the information theoretic approach is that the inclusion of the log of the prior odds of a match, as described by Howe and Lindsay (1981) and by Newcombe and Abbatt (1983) can be directly related to the methodology. Calculation of this test statistic yields a value which is commonly referred to as the "weight" for or against a match.

Given any pair of records, we want to make a decision as to whether they match (H_o -- the null hypothesis) or do not match (H_1 -- the alternative hypothesis). This decision will be based on the observed comparison of the attribute items on the two records. The set of all outcomes resulting from this comparison is the random variable, x_i, which takes values according to the outcomes which were specified for all of the attributes.

The discrete random variable, x_i, can take any of n different values. The number n can be very large, either because a large number of attributes are compared, or because a large number of outcomes are possible for any one attribute comparison. The probabilities with which x_i takes any of the n values under both H_o and H_1 are assumed to be known. The question of estimating these probabilities is addressed later. The decision process is formalized by considering the following two hypotheses:

H_o: The event that two records represent the same unit (i.e., a match). Under H_o, the frequency function of the random variable, x_i, is denoted $P(x_i/H_o) = p_{oi}$ for i=1, ... n.

H_1: The event that the two records represent different units (i.e., a nonmatch.) Under H_1, the frequency function of the random variable, x_i, is denoted $P(x_i/H_1) = p_{1i}$ for i=1, ... n.

AN EXAMPLE OF A COMPARISON VARIABLE

Assume that two records are being compared and that a decision will be made as to their match status based on a comparison of three attributes: surname, first name, and sex. For each attribute there will be two possible outcomes: either they agree or they do not agree. Thus, the comparison set can take any of 2**3 = 8 (n=8) possible values. For simplicity we also assume that the probabilities of agreement or disagreement of the attributes are independent under both H_o and H_1. Thus, given the following table of probabilities, the frequency function of the comparison vector can be calculated under both hypotheses.

TABLE I
PROBABILITIES OF AGREEMENT

Attribute	Under H_o	Under H_1
Surname	.90	.05
First name	.85	.10
Sex	.95	.45

In the following let $x=(a_1,a_2,a_3)$, where $a_i = 0$ if item i disagrees, and $a_i=1$ if item i agrees. The comparison of surname is represented by a_1, the comparison of first name by a_2, and the comparison of sex by a_3. Thus, the random variable, x_i, has the frequency functions given by p_{oi} (under H_o) and p_{1i} (under H_1) in the following table.

TABLE II
PROBABILITIES FOR COMPARISON VARIABLE

i	x_i	p_{oi}	p_{1i}
1	(0,0,0)	.0008	.4703
2	(1,0,0)	.0068	.0248
3	(0,1,0)	.0043	.0523
4	(0,0,1)	.0143	.3848
5	(1,1,0)	.0383	.0028
6	(1,0,1)	.1283	.0203
7	(0,1,1)	.0808	.0428
8	(1,1,1)	.7268	.0023

THE TEST STATISTIC

As shown in the Appendix, the test statistic

$$T(x_i) = \log(p_{oi}/p_{1i}) = I(o:1;x_i). \quad (1)$$

is a sufficient statistic for discriminating between H_o and H_1. The number log (p_{oi}/p_{1i}) is an information number. It provides a measure of the information for discriminating for H_o and against H_1 which was gained by observing the random variable, x_i.

$T(x_i)$ is the log of the ratio of the probability of the outcomes, denoted by x_i, under H_o to the probability of the same set of outcomes under H_1 (the log of the likelihood ratio.) Note that if these probabilities are the same then $T(x_i)=0$, and this set of outcomes has no discriminating power for identifying whether records represent the same unit. If p_{oi} is larger than p_{1i}, then $T(x_i)$ will be positive for that category. The larger $T(x_i)$, the stronger is the possibility that observation of this set of outcomes indicates that the records represent the same unit. If p_{oi} is smaller than p_{1i}, then $T(x_i)$ is negative. The smaller $T(x_i)$, the stronger is the possibility that this set of outcomes indicates that the records do not represent the same unit.

DETERMINING THE CRITICAL REGION

The final part of the matching problem is to determine cut-off values, c_1 and c_2, so that H_1 is rejected if $T(x_i)$ is greater than c_2 and H_o is rejected if $T(x_i)$ is less than c_1. If $T(x_i)$ falls between these two values, the test is inconclusive and the record pair may be subject to manual follow up.

In standard applications of testing simple hypotheses, there are only two outcomes: accept the null hypothesis or reject it. Here, the three region test comes from the union of two tests. First, consider a test of H_o vs. H_1. For a test with significance level alpha, this leads to the critical region defined by c_1. Next, consider the test of H_1 vs. H_o with significance level beta. This leads to a critical region defined by c_2. Individually, according to the Neyman-Pearson Lemma, these tests are the best tests at their respective significance levels. The first test rejects H_o if $T(x_i)$ is less than c_1. The second test rejects H_1 if $T(x_i)$ is greater than c_2.
Since c_1 is generally less than c_2, the union of these two tests yields the three region test described above.

This is illustrated below with our previous example. In Table III the column labeled $T(x_j)$ is the log of the ratio of p_{oj} and p_{1j} from Table II, but here the table is arranged so that the $T(x_j)$ are in ascending order. The next to last column presents the cumulative probability under H_o of observing $T(x_i)$ less than or equal to the given $T(x_j)$. It is used to specify c_1. In this example, if alpha is equal to .05, then c_1 is equal to -1.9. The last column is the cumulative probability under H_1 of observing $T(x_i)$ greater than or equal to the given $T(x_j)$. It is used to specify c_2. In this example, if beta is equal to .05 then c_2 is equal to 2.7.

TABLE III

THE DISTRIBUTION OF THE TEST STATISTIC

j	x_j	$T(x_j)$	p_{oj}	p_{1j}	$\sum_{k=1}^{j} p_{ok}$	$\sum_{k=j}^{n} p_{1k}$
1	(0,0,0)	-9.2	.0008	.4703	.0008	1.0004
2	(0,0,1)	-4.8	.0143	.3848	.0151	.5301
3	(0,1,0)	-3.6	.0043	.0523	.0194	.1453
4	(1,0,0)	-1.9	.0068	.0248	.0262	.0930
5	(0,1,1)	.9	.0808	.0428	.1070	.0682
6	(1,0,1)	2.7	.1283	.0203	.2353	.0254
7	(1,1,0)	3.8	.0383	.0028	.2736	.0051
8	(1,1,1)	8.3	.7268	.0023	1.0004	.0023

Thus, if alpha and beta both equal .05, we would classify a pair as a match if we observe vectors (1,0,1), (1,1,0), or (1,1,1). We would classify pairs as a nonmatch if we observe (0,0,0), (0,0,1), (0,1,0), or (1,0,0). If we observed (0,1,1): agreement on sex and first name, but disagreement on surname, we would be unable to classify the pair as either a match or a nonmatch.

The test statistic and critical region defined in this way are the same as those developed by Fellegi and Sunter (1969), although that paper also included a discussion of randomization to achieve the type 1 and type 2 error levels exactly. They develop the decision rule for accepting H_o or H_1 based on minimizing the probability of not making a decision. That is: minimizing the probability that $T(x_i)$ falls between c_1 and c_2 for a given alpha and beta.

THE POSTERIOR ODDS RATIO

The development presented here and in Fellegi-Sunter (1969) use the test statistic defined in equation (1). However, equation (A2) can be rewritten as

$$\log P(H_o/x_i)/P(H_1/x_i) = \log p_{oi}/p_{1i} + \log P(H_o)/P(H_1). \quad (2)$$

Here the log of the posterior odds ratio is written as the sum of the information number and the log of the prior odds ratio. Howe and Lindsay (1981) call equation (2) the "total weight" for a match, but acknowledge that the prior odds ratio is difficult to evaluate. The most recent papers by Newcombe and Smith include procedures for estimating the prior odds ratio in some unique situations (see Newcombe and Abbatt, 1983 and Smith, Newcombe, and Dewar, 1983). Note that the prior odds ratio reflects any information available regarding the match status of a given record pair before the attribute comparison. If the prior odds of a match were the same for each record pair then the test statistic and critical region for the comparison of attributes would both be shifted by the same value. In such a case the inclusion of the prior odds ratio would not change the outcome of the statistical test. However, the posterior odds ratio has the advantage that it can be interpreted directly as the odds that the record pair matches.

In the Smith, Newcombe, and Dewar paper, the prior odds ratio is calculated based on a life table analysis of the severity of cancer diagnosed, an attribute available in the search file, and the year of the death file being searched. In their example, the prior probability of a match is different for each individual in the search file and instead of applying specifically to a record pair, it applies to the individual record initiating the search and to an entire one year death file.

INDEPENDENCE OF ATTRIBUTES -- A SIMPLIFYING ASSUMPTION

In the original pages of this discussion, x_i was defined to be a discrete random variable which was the intersection of m attribute comparisons. If the result of each attribute comparison is denoted as t_j for j=1, ..., m, then x_i can be written as the intersection of the t_j:

$$x_i = t_1 \cap t_2 \cap \ldots \cap t_m.$$

If $t_1, \ldots, t_m$ are statistically independent, then equation (1) can be written as:

$$I(o:1;x_i) = \sum_{j=1}^{m} I(o:1;t_j).$$

Thus, if the set of attribute variables, t_j, are statistically independent, the weights (i.e., the information) for each t_j can be calculated separately, and the overall weight (the information contained in the intersection of the t_j) is just the sum of the weights for each t_j.

In the previous example, the three attributes were assumed to be independent. Hence, the weight for any observed vector can be calculated as the sum of the information associated with agreement or disagreement on each attribute. For example, for $x_i=(0,1,1)$ the weight can be calculated as the sum of the information associated with disagreement on surname,

$$T(a_1=0) = \log(.1/.95) = -3.25;$$

the information associated with agreement on first name,

$$T(a_2=1) = \log(.85/.1) = 3.09;$$

and the information associated with agreement on sex,

$$T(a_3=1) = \log(.95/.45) = 1.08.$$

The sum of these weights is .92, as shown in Table III for the weight (the value of $T(x_j)$) associated with the observation (0,1,1). Thus, if it is reasonable to assume that the outcomes of attribute comparisons for different attributes are statistically independent, then the calculation of the test statistic is simplified because the weights can be calculated separately and summed.

In this example, it is reasonable to assume that agreement on surname is independent of agreement on either first name or sex. However, if there is agreement on first name, it is likely that there will be agreement on sex. Hence, in this example, the assumption of independence does not really hold. To incorporate this dependence, one would need to consider the probabilities associated with the bivariate random variable.

AN EXAMPLE OF A MULTIPLE OUTCOME COMPARISON

The following is a vastly simplified example of defining the specific outcomes of attribute comparison by making use of the values they can assume. This type of "frequency" argument results in lower weights for agreement on common items and higher weights for agreement on rare items. It is a simplified version of the treatment of frequencies and error structures presented in the Fellegi-Sunter paper, pages 1192 and 1193 (pp. 60 and 61 in this volume).

Here, assume that surnames are being compared in a pair of records. Assume that there are only two frequently occurring names in the file, "Smith" and "Jones"; the other names (m of them) all occurring with roughly the same low frequency. Thus, we define the following set of outcomes of the comparison of surname:

$$x = \begin{cases} \text{"Smith"} & \text{if the two variables agree and both equal "Smith,"} \\ \text{"Jones"} & \text{if the two variables agree and both equal "Jones,"} \\ \text{"other"} & \text{if both variables agree but do not equal either "Smith" or "Jones,"} \\ \text{"disagree"} & \text{if the items disagree.} \end{cases}$$

(Note that the set of outcomes defined for item comparison must specify a partition of the set of all possible results into mutually exclusive and exhaustive subsets.)

Further assume that: 1) surnames in the two files under consideration are both random samples from the same population, and that in this population, "Smith" occurs with probability p_a, "Jones" occurs with probability p_b, and each of the other m error-free names in the file occurs with probability p_o; and 2) the only errors in the name fields are keypunch errors, which occur at the same rate, 1%, in both files, independent of the particular name.

Under H : A pair of records is a match. Names agree unless there is a keypunch error. Thus, the probability of agreement on Smith is $p_{o1} = p_a*(.99)**2$ (the probability of observing "Smith" times the probability that the value was keypunched correctly on both files). Similarly, the probability of agreement on Jones $p_{o2} = p_b*(.99)**2$, and the probability of agreement on one of the other names is $p_{o3}=p_o*(.99)**2$. The probability of disagreement on name when the record pairs represent the same individual is $p_{o4}= 1-p_{o1}-p_{o2}-m*p_{o3}$ $= (1-(.99)**2)*(p_a+p_b+m*p_o)$ $= 1-(.99)**2=.02$.

Under H_1: The records do not represent the same individual and any agreement on name occurs at random. The probability of agreement with name "Smith" is $(.99*p_a)**2$; the probability of agreement with name "Jones" is $(.99*p_b)**2$; the probability of agreement with some other name is $(.99*p_o)**2$; and the probability of disagreement on name is $1-.99**2*(p_a**2+p_b**2+m*p_o**2)$. (We have assumed that the probability that a keypunch error results in some valid name is negligible.)

Thus, from equation (1) the weight for the various outcomes is:

If x*=Smith,
$T(x*)=\log(.99**2*p_a/.99**2*p_a**2)=\log(1/p_a)$.

x*=Jones,
$T(x*)=\log(.99**2*p_b/.99**2*p_b**2)=\log(1/p_b)$.

x*=other,
$T(x*)=\log(.99**2*p_o/.99**2*p_o**2)=\log(1/p_o)$.

x*=disagree,
$T(x*)=\log$
$(.02/(1-*.99**2*(p_a**2+p_b**2+m*p_o**2)))$.

Newcombe, Kennedy, Axford, and James (1959) noted that in frequency based matching, if an item, a, is found in a master file with probability p_a, and if the two files being matched can be viewed as a sample from that master file, then, when a record pair is a match, the probability that the items agree and equal "a" is proportional to p_a. When the record pair is a nonmatch the probability is proportional to p_a**2 with the same constant of proportionality. Thus, the weight for a match when item a is observed is $\log(p_a/p_a**2) = \log(1/p_a)$. This is illustrated in the example above. Most of the Smith and Newcombe papers describe calculation of the weights for agreement on a particular item as the log of the inverse of the frequency of occurrence of that item.

The Fellegi-Sunter paper presents a derivation of the frequency based weights for specific agreement in the presence of several types of errors. Their procedure still leads to weights for agreement of $\log(1/p_a)$ because, as in the above example, the error terms impact the probability of agreement under H_o and the probability of agreement under H_1 in the same way.

VARIATIONS IN PRACTICE

Probabilistic matching techniques (based on the Fellegi-Sunter paper) have been implemented in many software systems, including the Generalized Iterative Record Linkage System (GIRLS) from Statistics Canada (see Smith and Silins, 1984) which is now called the Canadian Linkage System (CANLINK); UNIMATCH from the U.S. Bureau of the Census (see Jaro, 1972); the Statistical Reporting Service's (SRS) Record Linkage System from the U. S. Department of Agriculture (USDA); and the California Automated Mortality Linkage System (CAMLIS) from the University of California at San Francisco. Work by Rogot et al. (1983) at the National Center for Health Statistics has also used probabilistic matching techniques.

The two major references for this section are a paper by Howe and Lindsay (1981), which describes a version of the GIRLS system, and a number of unpublished papers by Richard Coulter, Max Arellano, William Arends, Billy Lynch, and James Mergerson dated 1976 and 1977, which describe the SRS Record Linkage System. These two systems were included in this review because they are applications of a modified Fellegi-Sunter approach and because the available documentation was thorough.

The GIRLS system was developed to support epidemiological research. Thus, it is primarily intended to link records for a cohort group to morbidity or mortality data. Attributes available for comparison usually include first name, surname, middle initial, sex, date of birth, place of birth, parents' names and places of birth. Some of the application-specific items, such as blocking attribute and definition of outcomes for attribute comparison, are not fixed in the system. They can be specified by the user. In the following, the specific applications by Howe and Lindsay are described.

The SRS record linkage system is intended to support development and maintenance of state-level sampling frames for agricultural surveys. Here, the primary intent of the linkage system is to unduplicate a list created by merging multiple lists. The most commonly available attributes are surname, first name, and address. In addition to the probabilistic matching procedure, record pairs which have identical address fields are reviewed manually to identify matches. This system is not a general-purpose matching system. It was developed and is used solely to maintain the USDA frames.

Blocking

In these applications, both systems block first on surname code -- a variation of the New York State Identification and Intelligence System (NYSIIS) code. A surname code is an alphabetic code designed so that the most similar names and the names with the most frequently encountered errors of misreporting will have the same code. See Lynch and Arends (1977) for a description of surname codes and the rationale used by SRS to select the NYSIIS code for their system. If the resultant block size is too big, SRS uses secondary blocking on first initial and tertiary blocking on location code. The Howe and Lindsay application blocks first on NYSIIS code, then on sex. In neither case are the weights changed to reflect the impact of blocking.

Weights for Agreement

Both systems make extensive use of frequency-based weights, and both systems use the files being matched to calculate the frequencies. Both systems also assume that these frequencies include keypunch errors, recording errors, and legitimate name changes. This is different from the Fellegi-Sunter approach, which assumed that the frequencies were based on an error-free name file.

The SRS approach handles partial agreements by calculating a weight for agreement on specific surname and a weight for agreement on specific NYSIIS code with disagreement on surname. The Howe-Lindsay paper extends the accounting for partial agreement by specifying agreement on specific first seven characters of surname; agreement on specific first four characters with disagreement on the next three characters; and agreement on specific NYSIIS code with disagreement on the first four characters of surname. In both systems, pairs with disagreement on NYSIIS code will never be considered because of the blocking.

Estimation of Error Rates

Both systems use an iteration scheme to provide final estimates for the required error rates. First, initial estimates are provided, a sample of records is processed through the matching algorithm, and a preliminary set of matched record pairs is identified. These pairs are assumed to be true matches and are used to estimate the error rates, as discussed below. These revised estimates for the error rates are input to the system; the sample is processed again and the newly matched pairs are used to reestimate the error rates. The iteration is continued until the estimates for the error rates converge.

The errors are handled in the Howe-Lindsay paper as transmission rates:

t_1 = the probability that the first seven characters of surname are equal to the "true" value;

t_2 = the probability that the first four characters are equal to the "true" value but the next three characters are different; and

t_3 = the probability that the surname code is equal to the "true" surname code, but that the surnames disagree in the first four characters.

These transmission rates can be estimated from a sufficiently large set of pairs which represent true matches by using the following counts: the number of pairs which agree on the first seven characters; the number of pairs which agree on the first four characters not on the next three, and the number which do not agree on the first four characters. The assumption is made that this set of matched pairs is representative of all possible matched pairs. Note that t_3 will be underestimated because of the blocking.

In the SRS system, the error rates used are:

e = the probability that a name is misreported or misrecorded

e_T = the probability that in a record pair which does represent the same unit, the names are correct but different.

These definitions of the error rates are the same as those used in the Fellegi-Sunter paper. The overall weights for specific agreement are different because the frequencies themselves are derived under different assumptions, as mentioned above. In the SRS system, the error rates are estimated from the set of pairs which represent true matches by using: the number of pairs which have the same name; the number which have different names; and the number which have similar names (where "similar" was not defined). Here, e_T will necessarily be underestimated because the blocking procedure assures that records will be compared only if they agree on NYSIIS code.

The Critical Region

Both systems use an empirical procedure to determine the critical region. That is, a frequency distribution of the weights for a sample of record pairs is plotted, and the critical values are selected based on the shape of the curve. As an alternative, the SRS system also calculates an initial lower critical region as the sum of the weights for agreement of the most common surname, first name, and location. The initial upper critical region is estimated as the initial lower critical region plus the weights for agreement on the most common middle name, route and box number. These calculated upper and lower regions are used during the iteration to estimate error rates. They are conservative since both are positive.

System Considerations

In the Howe-Lindsay approach, an initial blocking and comparison are done before the frequency based agreement weights are calculated. At this stage, only weights for disagreement are summed and as the accumulated weight becomes too negative, the record pair can be rejected as a possible match before all attributes have been compared. With this approach the order of adding in attributes is important, with those having the greatest negative weight for disagreement entering first. If the total disagreement weight is above the threshold, the record pair is a possible match. A separate file is created containing those possibly matched pairs. For each such pair, this file contains one record with the identification numbers of the two records, the results of the comparison of attributes, and the values taken (if needed for the weight calculation). This potential linked file is then sent to a separate subroutine for calculation of the weights.

Grouping

Both systems create groups consisting of all records which have been linked with each other. (Here linked means that the calculated test statistic is above the upper critical value.) As described in the Howe and Lindsay paper, the group is formed by first taking a single record and adding to the group any records which have been linked to it, then adding all records which were linked to those records, and so on. Additional subgroupings are considered when two records from different groups have a weight between the two critical values.

Interpretation of the groups depends on the application. In the SRS application, members of a group could all be duplicates to each other. In the SRS system, subgroups are analyzed manually. In some of the applications described by Howe and Lindsay, neither input file has any duplication, and there is at most one matched record for a given record in the search file. In this case the groups are analyzed to pick the pair which represents the most likely match, usually the pair with the highest weight.

SUMMARY

This paper has described the probabilistic matching procedures discussed by Fellegi and Sunter (1969) from an information theoretic point of view. This approach gives additional insight into the calculation of the posterior odds ratio as mentioned by Howe and Lindsay, and as implemented in the recent work of Newcombe and Smith. Additionally, it has described some of the differences between two of the major systems which have been implemented based on the Fellegi-Sunter paper. Major differences between systems are in accounting for partial matches, the definition of the error rates, and in the handling of groups of record pairs which are all linked to each other. The major differences between these systems and the Fellegi-Sunter approach are 1) that these systems base their frequency counts on files which are acknowledged to contain errors, and 2) that they use an empirical procedure to determine the critical region for the statistical test.

REFERENCES

Arellano, Max and Arends, William, "The Estimation of Component Error Probabilities for Record Linkage Purposes," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished, May 1976.

Arellano, Max, "Optimum Utilization of the Social Security Number for Matching Purposes," "Weight Calculation for the Place Name Comparison," "Calculation of Weights for Partitioned Variable Comparisons (Trailing Blanks Case)," "Estimation of Component Error Probabilities for Record Linkage Purposes," "Development of A Linkage Rule for Unduplicating Agricultural Lists," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished papers, 1976 and 1977.

Arellano, Max and Coulter, Richard, "Weight Calculation for the Given Name Comparison," "Weight Calculation for the Middle Name Comparison," "Weight Calculation for the Surname Comparison," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished papers, 1976 and 1977.

Coulter, Richard, "An Application of a Theory for Record Linkage," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished, March 1977.

Coulter, Richard and Mergerson, James, "An Application of a Record Linkage Theory in Construction of a List Sampling Frame," Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., April 1977.

Fellegi, Ivan and Sunter, Alan, "A Theory for Record Linkage," Journal of the American Statistical Association, 1969, pp. 1183-1210.

Howe, G. R. and Lindsay, J., "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies," Computers and Biomedical Research 14, Academic Press, 1981, pp. 327-340.

Jaro, Matthew, "UNIMATCH--A Computer System for Generalized Record Linkage Under Conditions of Uncertainty," AFIPS Conference Proceedings, Vol. 40 for Spring Joint Computer Conference, May 1972, pp. 523-530.

Jaro, Matthew, "UNIMATCH--Generalized Record Linkage Applied to Urban Data Files," Proceedings of the American Statistical Association.

Kelley, Patrick, "A Preliminary Study of the Error Structure of Statistical Matching," Proceedings of the American Statistical Association, Social Statistics Section, 1983.

Kelley, Patrick, "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," Record Linkage Techniques--1985, Internal Revenue Service.

Kullback, Soloman, Information Theory and Statistics, Dover Publications, Inc., New York, New York, copyright 1968.

Lynch, Billy and Arends, Williams, "Selection of a Surname Coding Procedure for the SRS Record Linkage System," Sample Survey Research Branch, Research Division, Statistical Reporting Service, U. S. Department of Agriculture, Feb 1977.

Newcombe, Howard, "Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories," American Journal of Human Genetics, Vol 19, No. 3, Part I, (May) 1967.

Newcombe, Howard, and Kennedy, James, "Record Linkage: Making Maximum Use of Discriminating Power of Identifying Information," Communications of the Association for Computing Machinery 5, 1962, pp. 563-566.

Newcombe, H., Kennedy, J., Axford, S., and James, A.,"Automatic Linkage of Vital Records," Science, 130, 1959, pp. 954-959.

Newcombe, H., and Abbatt, J., "Probabilistic Record Linkage in Epidemiology," Report Prepared for Eldorado Resources, Ltd., Oct. 1983.

Rogot, Eugene, Schwartz, Sidney, O'Connor, Karen and Olsen, Christina, "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," Proceedings of the American Statistical Association, Section on Business and Economic Statistics, 1983.

Smith, Martha, Newcombe, Howard, and Dewar, Ron, "Proposed Procedure for the Alberta Cancer Registry Death Clearance," Health Division, Statistics Canada, (OEHRU-No. 1), March 1983.

Smith, Martha, Newcombe, Howard, and Dewar, Ron, "The Use of Diagnosis in Cancer Registry Death Clearance," Health Division, Statistics Canada, (OEHRU-No. 2), April 1983.

Smith, Martha and Silins, John, "Generalized Iterative Record Linkage System," (An excerpt), Statistical Uses of Administrative Records: Recent Research and Present Prospects, Department of the Treasury, Internal Revenue Service, July 1984.

U. S. Department of Commerce, Office of Federal Statistical Policy and Standards, Report on Exact and Statistical Matching Techniques, Statistical Policy Working Paper 5, 1980.

APPENDIX

This appendix presents a derivation of the test statistic for determining whether a record pair is a match or a nonmatch using an information theoretic approach (see Kullback, 1968).

WHAT IS AN INFORMATION NUMBER?

Given the prior probabilities associated with a match and a nonmatch, $P(H_o)$ and $P(H_1)$, we use Bayes theorem to calculate the posterior probabilities of H_o and H_1 based on the observed attribute comparison, x_i:

$$P(H_o/x_i) = P(H_o)*p_{oi}/(P(H_o)*p_{oi} + P(H_1)*p_{1i})$$

$$P(H_1/x_i) = P(H_1)*p_{1i}/(P(H_o)*p_{oi} + P(H_1)*p_{1i}).$$

Dividing these gives the posterior odds ratio:

$$P(H_o/x_i)/P(H_1/x_i) = P(H_o)*p_{oi}/(P(H_1)*p_{1i}),$$

and taking the logarithm (to any base) gives:

$$\log P(H_o/x_i)/P(H_1/x_i) = \log p_{oi}/p_{1i} + \log P(H_o)/P(H_1). \quad (A1)$$

This is the log of the posterior odds ratio or equivalently, the log of the posterior likelihood ratio. It can be rearranged to get:

$$\log p_{oi}/p_{1i} = \log P(H_o/x_i)/P(H_1/x_i) - \log P(H_o)/P(H_1). \quad (A2)$$

This number is the difference between the log of the posterior odds ratio and the log of the prior odds ratio. Thus, it provides a measure of the information for discriminating in favor of H_o against H_1 which was gained by observing the random variable x_i.

For this reason, the information gained by the set of outcomes of the attribute comparison, x_i, is defined to be:

$$I(o:1;x_i) = \log p_{oi}/p_{1i}. \quad (A3)$$

THE MEAN INFORMATION

The mean information for discriminating in favor of H_o against H_1 is the expected value of $I(o:1;x_i)$ under H_o, or

$$\begin{aligned} I(0:1) &= E_o(\log p_{oi}/p_{1i}) \\ &= \sum_{i=1}^{n} p_{oi} * \log p_{oi}/p_{1i}. \quad (A4) \end{aligned}$$

Here E_o represents the expectation under H_o. Note that the mean information is simply the expected value of the log of the likelihood ratio under H_o.

One useful mathematical fact is that I(o:1) is always greater than or equal to zero, with equality only when $p_{oi} = p_{1i}$ for all i = 1, ..., n. This gives an approach to selecting between the two hypotheses. Given any sample, it is possible to evaluate the sampling distribution under both hypotheses, and to calculate the mean information between the sampling distribution and the hypothesized distribution. The hypothesized distribution which was closer to the sampling distribution, as measured by the mean information, would be preferred.

THE TEST STATISTIC

When we compare the attributes associated with any two records, the result is one of the n possible values taken by x_i. We denote this observed random variable as x^*. The probability of observing $x^*=x_i$ is p_{oi} under H_o and p_{1i} under H_1. Thus, the sampling distribution of x* is simply;

$p_i=1$ if $x^* = x_i$, $p_i=0$ if x^* ne x_i.

We can write the mean information between the sampling distribution and H_o as

$$I(x^*:H_o) = \log(1/p_{oi}) \text{ for } x^*=x_i,$$

and the mean *information* between the *sampling* distribution and H_1 as

$$I(x^*:H_1) = \log(1/p_{1i}) \text{ for } x^*=x_i.$$

The decision rule, as described in Kullback (1968, chapter 5), is to pick the hypothesis which has the smallest mean information relative to the sampling distribution. That is, we accept the hypothesized distribution which is closest to the sampling distribution.

Thus, the procedure would be to accept H_o if $I(x^*:H_1)-I(x^*:H_o)$ is positive (or "sufficiently large.") and accept H_1 if it is negative (or "sufficiently small.")

This yields the test statistic, T(x*), where

$$T(x^*) = I(x^*:H_1)-I(x^*:H_o)$$
$$= \log(p_{oi}/p_{1i}) \text{ for } x^*=x_i. \qquad (A5)$$

T(x*) is the log of the *ratio* of the probability of the set of outcomes, x*, under H_o to the probability of x* under H_1. Note that if these probabilities are the same then T(x*)=0, and this set of outcomes has no discriminating power for identifying whether records represent the same unit. If p_{oi} is larger than p_{1i}, then T(x*) will be positive for that category. The larger T(x*), the stronger is the possibility that observation of this set of outcomes indicates that the records represent the same unit. If p_{oi} is smaller than p_{1i}, then T(x*) is negative. The smaller T(x*), the stronger is the possibility that this set of outcomes indicates that the records do not represent the same unit.

Since $T(x^*) = \log(p_{oi}/p_{1i})$ with probability p_{oi} under H_o, and with probability p_{1i} under H_1, the ratio of the probability that $x^*=x_i$ and the probability that $T(x^*) = T(x_i)$ is equal to 1.

Since the ratio of the probability function of x_i and the probability function of $T(x_i)$ does not depend on the p_{oi} or p_{1i}, $T(x_i)$ is a sufficient statistic for discriminating between H_o and H_1.

ADVANCES IN RECORD LINKAGE METHODOLOGY: A METHOD FOR DETERMINING THE BEST BLOCKING STRATEGY

R. Patrick Kelley, Bureau of the Census

I. INTRODUCTION

The term record linkage, as it will be used in this paper, is a generic term for any process by which the set of reporting units common to two or more files of data is determined.

Historically, government agencies have been the primary users of record linkage techniques. The reasons such agencies carry out record linkage projects are as varied as the purpose and scope of the agencies themselves. Consider the following examples:

a) The United States Department of Agriculture uses record linkage to update mailing lists (see Coulter and Mergerson, 1977).

b) Statistics Canada uses record linkage as a tool in epidemological research (see Smith, 1982).

c) The United States Census Bureau uses record linkage as a tool in coverage and content evaluation (see Bailar, 1983).

For a more detailed discussion of the history and and use of record linkage by United States government agencies see U.S. Department of Commerce (1980).

As an area of study, Record Linkage, with its associated statistical problems, is a special case of a larger area of concern. This area makes use of various mathematical and statistical techniques to study the problems involved in the classification of observed phenomena.

Discriminant analysis, discrete discriminant analysis, pattern recognition, cluster analysis and mathematical taxonomy are some of the specific fields which study various aspects of the classification problem. While record linkage contains its own specific set of problems it also has a great deal in common with these other fields.

The basic unit of study in the linking of two files F1 and F2 is F1XF2, the set of ordered pairs from F1 and F2. Given F1XF2, our job is to classify each pair as either matched or unmatched. This decision will be based on measurements taken on the record pairs. For example, if we are linking person records, a possible measurement would be to compare surnames on the two records, and assign the value 1 for those pairs where there is agreement and 0 for those pairs where there is disagreement. These measurements will yield a vector, Γ, of observations on each record pair.

The key fact which will allow us to link the two files is that Γ behaves differently for matched and unmatched pairs. Statistically we model this by assuming that Γ is a random vector generated by $P(\cdot \mid M)$ on matched pairs and $P(\cdot \mid U)$ on unmatched pairs. Thus, the Γ value for a single randomly selected record pair is generated by $pP(\cdot \mid M)+(1-p)\,P(\cdot \mid U)$ where p is the proportion of matched records.

This model for the record linkage problem is the same as the one used in discriminant analysis. In particular, as Γ is almost always discrete, the literature on discrete discriminant analysis is extremely useful (see for example Goldstein and Dillon, 1978). There are, however, several areas of concern that seem to be a great deal more important for record linkage than for the other classification techniques.

Our topic of discussion in this paper, blocking, arises from consideration of one of these problem areas. That area concerns the extreme size of the data sets involved for even a relatively small record linkage project. The size problem precludes our being able to study all possible record pairs. So, we must determine some rule which will automatically remove a large portion of record pairs from consideration. Such a rule is referred to as a blocking scheme since the resulting subset of record pairs often forms rectangular blocks in F1XF2.

The literature on the blocking problem is not extensive. Brounstein (1969), Coulter and Mergerson (1977) and U.S. Department of Commerce (1977) contain discussions of the practical aspects of choosing a blocking scheme; however, they provide no general framework within which to make such a selection. Jaro (1972) provides a framework for the selection of a blocking scheme but doesn't discuss the errors induced by blocking. Many other papers, particularly those on clerical matching, contain implicit information on blocking. But so far there has been no systematic study of this area.

To provide such a study we begin with the following three questions:

1) What are the benefits and costs involved in blocking and how do we measure them?

2) How do we select between competing blocking schemes? Is there a best scheme?

3) How do the various computing restrictions effect our blocking scheme selection?

These three questions will serve as a guideline for our investigation of the blocking problem. But, before we begin this investigation, we need to consider some background material on record linkage.

II. BACKGROUND

Again, our job in linking the two files F1 and F2 is to classify each record pair as either matched or unmatched. In practice, however, we usually include a clerical review decision for tricky cases. So, our set of possible decisions is

A1: the pair is a match
A2: no determination made - clerical review
A3: the pair is not a match.

Now, consider the class of decision functions $D(\cdot)$ which transform our space of comparison vector values, elements of which we will denote by γ, to the set of decisions $\{A1,A2,A3\}$. Given two or more decision functions in this class, what criterion will we use to choose between them?

In Fellegi and Sunter (1969) the argument is put forward that, as decision A2 will require costly and error prone clerical review, we should pick a decision procedure which will minimize the expected number of A2 decisions while keeping a bound on the expected number of pairs which are classified in error. Since the unconditional distribution of the comparison vector is the same for any randomly chosen pair, this reduces to picking that decision procedure which will minimize $P(A2)$ subject to $P(A1|U) <= \mu$ and $P(A3|M) <= \lambda$.

Given that you know $P(\cdot\,|M)$ and $P(\cdot\,|U)$, Fellegi and Sunter prove that the decision procedure which solves this problem is of the form

$$D(\gamma) = \begin{cases} A3 \text{ if } \ell(\gamma) <= t1 \\ A2 \text{ if } t1 < \ell(\gamma) < t2 \\ A1 \text{ if } \ell(\gamma) >= t2 \end{cases} \tag{1}$$

where $\ell(\gamma) = P(\gamma\,|M)/P(\gamma\,|U)$, t1 is the largest value in the range of $\ell(\cdot)$ for which $P(A3|M) <= \lambda$ and t2 is the smallest value in the range of $\ell(\cdot)$ for which $P(A1/U) <= \mu$.

It is this decision procedure that forms the basis for our study of the blocking problem.

III. MEASUREMENT OF THE COST AND BENEFIT OF BLOCKING

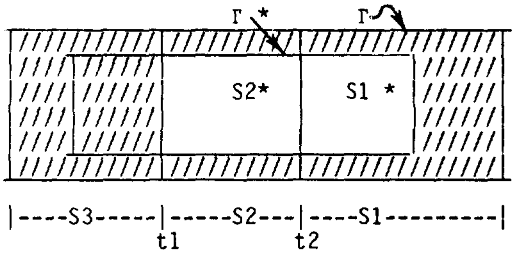
In the past sections we have outlined the more general aspects of record linkage and defined the blocking problem. In this section we will discuss blocking in the context of the decision procedure given in section II.

We base our general blocking strategy on the fact that the proportion of matched pairs in F1XF2 is small. So we will concentrate on blocking rules in which the pairs removed by the rule will be assigned the status of unmatched.

Fellegi-Sunter (1969) provides a formal model for blocking. This model defines a blocking scheme to be a subspace, say Γ^*, of the comparison space. Kelley (1984) provides a preliminary study of selected methods of measuring cost and benefit. The method found to have the most intuitive appeal is one that is based on the following amended decision procedure:

$$D'(\gamma) = \begin{cases} A3 \text{ if } \ell(\gamma) <= t1 \text{ or } \gamma \in \Gamma^{*C} \\ A2 \text{ if } t1 < \ell(\gamma) < t2 \text{ and } \gamma \in \Gamma^* \\ A1 \text{ if } \ell(\gamma) >= t2 \text{ and } \gamma \in \Gamma^* \end{cases} \tag{2}$$

A Venn diagram of this situation is given by



where S3* is represented by the shaded region. In this design Si and Si* are the regions of Γ values for which we make decision Ai under decision functions given by (1) and (2), respectively.

The error levels for this amended decision rule are given by

$$P(S3^* \mid M) = P(S3 \mid M) + P(S3^* - S3 \mid M)$$

$$= \lambda + P(S3^* - S3 \mid M).$$

and

$$P(S1^* \mid U) = P(S1 \mid U) - P(S1 \cap S3^* \mid U)$$

$$= \mu - P(S1 \cap S3^* \mid U).$$

These equations give us a means to compute a cost incurred by blocking on the subspace Γ^*, namely, $P(S3^* - S3 \mid M)$, the increase in probability of a false nonmatch. The benefit gained from blocking on Γ^* takes the form of a decrease in the number of pairs which will have to be processed. We will measure this benefit by the unconditional probability that a randomly chosen record pair yields a Γ vector in the block.

Now, given two blocking schemes which both have cost less than or equal to a fixed amount, the preferred scheme is the one with greatest benefit. Thus, we define the best blocking scheme to be that scheme which minimizes $P(\Gamma^*)$ subject to $P(S3^*-S3|M) <= w$, where w is an independently determined upper bound on blocking costs.

IV. COMPUTING THE BEST BLOCKING SCHEME - THE ADMISSIBILITY CONCEPT

Since the comparison vector is discrete, the computation of the best blocking scheme will require a comparison of all competing schemes. So, it's in our best interest to reduce the number of competing schemes. To make this reduction we note that if $\Gamma1^*$ and $\Gamma2^*$ are two competing schemes such that $\Gamma1^*$ is a subset of $\Gamma2^*$ then $\Gamma1^*$ is uniformly better than $\Gamma2^*$. So, we can remove $\Gamma2^*$ from the set of competing blocking schemes. The following definition formalizes this example:

> Γ^* will be said to be an admissible blocking scheme at $w = w0$ if
> a) $P(S3^* - S3 \mid M) <= w0$ and
> b) for every Γ^{**} that is a subset of Γ^*
> $P(S3^{**} - S3 \mid M) > w0$.

The concept of an admissible blocking scheme given by this definition is analogous to the concept of an admissible decision procedure. It serves to reduce, hopefully to a reasonable size, the number of blocking schemes competing for best. But, unfortunately, when actually applied to the task of computing the set of admissible blocking schemes, this definition is very cumbersome. The following lemma gives necessary and sufficient conditions for admissibility which are more favorable to algorithm development:

Lemma 1:

Γ^* is admissible at $w = w0$ if and only if
$\Gamma^* \cap S3 = \emptyset$ and
$P(\gamma|M) > w0 - P(S3^*-S3|M) \geq 0$ for all γ in Γ^*.

Proof:

If Γ^* is admissible then $P(S3^*-S3|M) <= w0$. Further, for $\Gamma^{**}= \Gamma^* - \{\gamma\}$ we have $P(S3^{**} - S3|M) > w0$. But $S3^{**} - S3 = (S3^* - S3) \cup (\{\gamma\}-S3)$. So, $P(\{\gamma\}-S3|M) + P(S3^* - S3|M) > w0$.

From this relationship we see that if γ is in S3 then $P(S3^*-S3|M) > w0$; thus, $\Gamma^* \cap S3 = \emptyset$. So we have $P(\gamma|M) > w0 - P(S3^*- S3|M)$ for all γ in Γ^*.

Conversely, we first note that $P\ (S3^*-S3|M) <= w0$. Next, let Γ' be a proper subset of Γ^* then Γ' is a subset of $\Gamma^*- \{\gamma\}$ for some γ. So, $P(S3'-S3|M)> = P(S3^*-S3|M) + P(\{\gamma\}-S3|M)$. Thus, we have $P(S3'-S3|M) >= P(S3^*-S3|M) + P(\gamma|M) > w0$. Hence, Γ^* is admissible.

Now, in theory, we can use the result of lemma 1 to compute all admissible schemes. However, since the minimum number of dimensional Γ vector values is 2**n, we would have to generate and classify on the order of 2**(2**n) subsets.

For n=5 this yields 4,294,967,300 subsets, which is clearly too large for practical consideration. So, while the admissibility concept is helpful in reducing the number of competing schemes, it hasn't served to provide us with a practical algorithm for the computation of the best blocking scheme. In the next section, we will give more attention to the development of such an algorithm.

V. IMPLEMENTATION CONSIDERATIONS

The previous section provides a general framework for studying blocking; however, it doesn't give us much insight into the practical side of determining a block of records for possible linkage. If we keep in mind that I/O and computing the comparison vector are the biggest consumers of time in the linkage operation we see that admissible blocking schemes that require the computation of a Γ vector value for each record pair are not practical. Thus, though a scheme might be theoretically admissible it might not be feasible.

One solution for this problem is to block by using certain fields on the record (such as soundex code of surname or address range) as sort keys. The blocks would be determined by those record pairs with equal keys. Thus, the match status of unmatched pairs would be implicitly assigned to all record pairs with unequal keys.

Restricting our study to blocking schemes which are determined by sort keys implies that the comparison vector we want to use will consist of dichotomous components measuring agreement on the record identifier fields. We will further assume that the components of the comparison vector are stochastically independent for both matched and unmatched record pairs.

Now, letting $mi = P(\Gamma i=1|M)$, $ui=P(\Gamma i=1|U)$ and Γ^* be the blocking scheme determined by sorting on components i1,...,ik we have the following result:

Lemma 2:

Suppose that mi>1/2 and ui<mi for all i then Γ^* is admissible at w0 if and only if

a) $w0 - P(S3^*-S3|M) >= 0$
b) $P(\gamma^*|M) > \mathrm{Max}\ \{t1P(\ \gamma^*|U), w0 - P(S3^*-S3|M)\}$,

where γ^* is such that $\gamma i1^* = 1,\ldots, \gamma ik^* = 1, \gamma ik{+}1^* = 0, \ldots, \gamma ip^* = 0$.

Proof:

First suppose that Γ^* is admissible at w0 then conditions a) and b) follow directly from lemma 1 and the fact that $P(\gamma|M) > t1\ P(\gamma|U)$ for all γ in $S3^c$.

Now, to establish the converse we first note that, since $mi > 1/2$ for all i, $P(\gamma^*|M) = \min_{\gamma\varepsilon\Gamma^*} P(\gamma|M)$. So $P(\gamma|M) > w0 - P(S3^*-S3|M) >= 0$ for all γ in Γ^*. Next we need to prove that $\Gamma^* \cap S3 = \emptyset$. To prove this we note that $ui < mi$ implies that $mi/ui > (1-mi)/(1-ui)$. So, $P(\gamma|M)/P(\gamma|U) > P(\gamma^*|M)/P(\gamma^*|U)$ for all γ in Γ^*. Thus, $\Gamma^* \cap S3 = \emptyset$. The converse follows from lemma 1.

In comparing lemma 2 with lemma 1, we see that lemma 2 has a definite computational advantage above and beyond the reduction in competing schemes gained by restricting attention to those schemes based on sorting. That advantage lies in the requirement to check for admissibility at only one point in the blocking scheme, namely γ^*. This results in tremendous savings in computing time and simplifies algorithm construction and coding considerably. In the next section we apply lemma 2 to a simple numeric example.

VI. AN EXAMPLE

As an example, let's consider matching two files of records based on the identifiers surname, first name, and sex.

Suppose we have determined beforehand that,

for surname	m1 = .90	and u1 = .05,
for first name	m2 = .85	and u2 = .10,
and for sex	m3 = .95	and u3 = .45.

Retaining the assumption of the previous section our discriminant function is given by

$$L(\gamma)= \ln 2(l(\gamma)) = \sum_{i=1}^{3} [\gamma i\ \ln 2\ (mi/ui) +(1-\gamma i)\ \ln 2\ ((1-mi)/(1-ui))].$$

To compute the Fellegi-Sunter decision procedure we first compute L for each agreement pattern and then we order the patterns on increasing L. The following table gives the results of this operation:

Pattern	Sum of P(·|M)	One minus sum of P(·|U)	L
(0,0,0)	.00075	.52975	-9.29
(0,0,1)	.01500	.14500	-4.76
(0,1,0)	.01925	.09275	-3.62
(1,0,0)	.02600	.06800	-1.87
(0,1,1)	.10675	.02525	.92
(1,0,1)	.23500	.00500	2.67
(1,1,0)	.27325	.00225	3.79
(1,1,1)	1.00000	0.00000	8.34

Using this table it is clear how one would compute t1 and t2 for given λ and μ .

For example, if we let λ = .05 and μ = .05 then t1 = -1.87 and t2 = 2.67. The actual values of λ and u are .026 and .02525, respectively. We will use this decision procedure to discuss the blocking problem.

Consider our space of admissible blocking schemes based on sorting. We note that since no single component blocking scheme is admissible, we have a total of four schemes to test. Now, for convenience let B1 denote blocking on surname and first name, B2 denote blocking on surname and sex, B3 denote blocking on first name and sex, and B4 denote blocking on all components.

The following table gives the information necessary to determine the admissibility of Bi:

Bi	P(S3*-S3|M)	P(γ*|M)	values of w0 for which Bi is admissible
B1	.209	.03825	.209 ≤ w0 < .24725
B2	.119	.12825	.119 ≤ w0 < .24725
B3	.1665	.08075	.1665 ≤ w0 < .24725
B4	.24725	.72675	.24725 ≤ w0 < .974

Before we go on it is interesting to note that the minimum w0 value for which any of the B_i is admissible is .119. Thus, the minimum loss we can incur by blocking is an increase in false non-match probability of .119.

Looking at the admissible blocking schemes as a function of w0, we have the following:

1. For .119 ≤ w0 < .1665 B2 is admissible.
2. For .1665 ≤ w0 < .209 B2 and B3 are admissible.
3. For .209 ≤ w0 < .24725 B1, B2, B3 are admissible.
4. For .24725 ≤ w0 < .974 B4 is admissible.

Now, to compute the best admissible blocking scheme we must determine which of the competing schemes has the smallest probability of occurrence. The probability of occurrence of schemes Bi, say P(Bi), is given by pP(Bi|M)+(1-p)P(Bi|u), where p is the proportion of matched record pairs. Thus, in general, the best admissible scheme will be a function of p.

To compute the best blocking scheme for cases 2 and 3 consider the following table:

	P(Bi|M)	P(Bi|U)
B1	.765	.005
B2	.855	.0225
B3	.8075	.045

So, for case 2, B2 is the best blocking scheme for values of p <= .3214 and B3 is the best blocking scheme for p > .3214. For case 3, B1 is uniformly the best blocking scheme.

At this point, we have demonstrated how to select the best blocking scheme for a fixed value of w0. But it still is unclear how one would use this information to actually make a decision about which scheme to use. To study this question let's consider the nature of such a decision. To select a blocking scheme we need to balance the cost with the overall benefit. Let's redo our example this time for several different values of w0 and compare the benefits for the resulting schemes.

The following is the first part of the list of the best blocking schemes for all values of w0. This list is presented in increasing order of w0. The expected benefit, in terms of the percent of F1XF2 that would be examined, is given for each scheme. To compute this benefit the approximate sizes of F1 and F2 are required. We used F1 size = 200,000 and F2 size = 100,000 in this example.

1. Admissible blocking schemes at w0=0.0492501 are as follows:
   The scheme determined by sorting on sex.
   The expected percent of the cross product of this blocking scheme would examine is bounded above by 45.00005%.
2. Admissible blocking schemes at w0=0.0992500 are as follows:
   The scheme determined by sorting on surname.
   The expected percent of the cross product this blocking scheme would examine is bounded above by 5.00009%.
3. Admissible blocking schemes at w0=0.1442501 are as follows:
   The scheme determined by sorting on surname and sex.
   The expected percent of the cross product this blocking scheme would examine is bounded above by 2.25008%.
4. Admissible blocking schemes at w0=0.149250 are as follows:
   The scheme determined by sorting on first name.
   The scheme determined by sorting on surname and sex.
   Of these, the best blocking strategy, as a function of the proportion of matched pairs, is as follows:

   For p=0.000000000 to p=0.939394700 sort on components surname and sex.
   For p=0.939394700 to p=1.000000000 sort on components first name.
   The expected percent of the cross product this blocking scheme would examine is bounded above by 2.25008%.

To use this list for decision-making purposes one would have to have some idea about how much data they can afford to look at and how large a false non-match rate they could tolerate. For example, in looking at the scheme determined by sorting on sex, we have a small (though maybe not small enough) w0 value but the number of record pairs we would have to look at would be around 9x10**10, which is clearly not feasible. Sorting on surname has a slightly higher w0 value, but reduces the number of records to 10**10. If we are willing to accept an even higher w0, then we can sort on surname and sex, which further reduces the number of record pairs to 4.5x10**9.

Another important piece of information that we shouldn't overlook is the number of record pairs we can hold in memory at any one time. We don't want to select a blocking scheme for which the individual block sizes are too large. So not only is the total number of pairs in the block important but so is the number of states of the sorting variable and the distribution of that variable over those states.

VII. SUMMARY

The blocking problem is intrinsic to record linkage. As such, before a link between files is attempted a decision must be made concerning the appropriate blocking method.

In this paper we study this decision, along with its costs and benefits, through the record linkage methodology developed in Fellegi and Sunter (1969). This methodology applies classic decision theory techniques to the record linkage problem, constructing the optimum classifer under a loss function analogous to that of hypothesis testing.

The result of our study is a method which can be used to balance the cost and benefit of blocking. This method involves maximizing benefit subject to an upper bound on cost. The measurement of cost and benefit is based on the Fellegi-Sunter method and, as such, makes use of a similar loss function.

NOTES AND REFERENCES

Bailar, Barbara A. (1983), Counting or Estimation in a Census -- A Difficult Decision, Proceedings of the American Statistical Association, Social Statistics Section, pp. 42-49.

Brounstein, S. H. (1969), Data Record Linkage Under Conditions of Uncertainty, delivered at the Seventh Annual Conference of the Urban and Regional Information Systems Association.

Coulter, Richard W. and Mergerson, James, W. (1977), An Application of a Record Linkage Theory in Constructing a List Sampling Frame. List Sampling Frame Section, Sample Survey Research Branch, Statistical Reporting Service, U.S. Department of Agriculture.

Fellegi, Ivan and Sunter, Alan (1969), A Theory for Record Linkage, Journal of the American Statistical Association, vol. 64, pp. 1183-1210.

Goldstein, Matthew and Dillon, William (1978), Discrete Discriminant Analysis, Wiley.

Jaro, M. A. (1972), UNIMATCH - A Computer System for Generalized Record Linkage Under Conditions of Uncertainty, AFIPS - Conference Proceedings, vol. 40, pp. 523-530.

Kelley, Robert Patrick (1984), Blocking Consideration for Record Linkage Under Conditions of Uncertainty, Statistics of Income and Related Administrative Record Research: 1984, Department of the Treasury, Internal Revenue Service, pp. 163-165.

Smith, Martha E. (1982), Value of Record Linkage Studies in Identifying Population at Genetic Risk and Relating Risk to Exposures. Progress in Mutation Research, vol. 3, pp. 85-98.

U. S. Department of Commerce, National Bureau of Standards (1977), Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers.

U.S. Department of Commerce, Office of Federal Statistical Policy and Standards (1980), Statistical Policy Working Paper 5 - Report on Exact and Statistical Matching Techniques.

DISCUSSION

Eli S. Marks, Consultant

WINKLER

This paper discusses Bill Winkler's presentation on "Preprocessing of Lists and String Comparison."

Key factors in "Preprocessinq of Lists" are:

1. The objectives of the system and the costs of various levels and types of matching error.
2. Costs of attaining a given matching accuracy level by preprocessing vs. other alternatives (e.g., suitably tailored "tolerances").
3. The nature of the matching system--manual, computerized, "mixed," etc.
4. How preprocessing is performed.

1. Objectives

The objectives of the system and the costs of matching error are intimately related. For example, if the objective is to estimate undercoverage of the U.S. census in each state, city, county, township, place, etc. for purposes of allocation of representation in Congress and state legislatures, city/county councils, etc. and for allocating federal and state funds to state and local jurisdictions, a uniform level of matching error everywhere is more important than the absolute level of matching error. Thus, preprocessing may have little value if its effect is to reduce the different types of matching errors by the same percentages in all jurisdictions. On the other hand, if preprocessing reduces urban matching error more than rural, it may be desirable or undesirable, depending upon whether the level of urban matching error without preprocessing is greater or less than the level of rural matching error without preprocessing.

2. Alternative Techniques

The objective of preprocessing (i.e., reduction of matching errors) can be attained by other means (e.g., the prescription of matching "tolerances"); and these techniques may cost less than preprocessing. For example, soundex coding is a form of "matching tolerance." That is, all disagreements of vowels and some disagreements of consonants are ignored in determining whether a pair of records match on the soundexed "identifier." One can, in fact, combine some preprocessing with tolerances (and, perhaps, other error-reducing techniques) to get a more efficient matching system than either can give alone. For example, one can prescribe standard abbreviations for the address suffixes "Avenue," "Street," "Road," "Drive," "Place," "Boulevard," etc., but also provide that an address match where the suffixes differ will be accepted unless there is another address match where the suffixes agree. For example, "Sutton Drive" would match "Sutton Road" unless either file contains both "Sutton Road" and "Sutton Drive."

Standard spelling of name and address may be achieved more accurately and more cheaply by controlling data collection, recording and "keying" (to put the data in machine readable form) than by preprocessing. This would, for example, avoid most of the errors of preprocessing by ZIPSTAN exhibited by the examples shown in the paper. Preprocessing errors can also be reduced or eliminated by other means, such as the clerical insertion of distinctive symbols to designate components of name and address, as outlined in Section 4 below.

It should be noted that selection of an "optimum matching strategy" is heavily dependent upon the type(s) of matching system(s) considered and that the choice of type of matching system is a vital part of the determination of "optimum matching strategy."

3. Kind of Matching System

The paper by Winkler notes that matching systems can be manual or computerized and implies that preprocessing is largely unnecessary for manual matching systems. I think his suggestion that individuals can usually determine accurately whether a pair of name and address records is actually a match or nonmatch is somewhat optimistic. Individuals can make this determination (so can a computer system), but how accurately depends on the kind of system. The great advantage of a competent human matcher operating in a properly designed matching system is the use of judgmental flexibility, provided, of course, he or she has good judgment and the matching rules permit him (her) to use that judgment (and I have seen many sets of matching instructions which do not). The great disadvantage of a well-designed manual matching system with competent matchers is the human matcher's slowness and the inevitable drop in efficiency in operating in a system which requires examining large masses of records; and not in lack of clear decision rules, inconsistency of application of decision rules, and nonreproducibility of results. All of the latter do occur, but can be adequately controlled in a well-designed matching system (although it is not easy!). However, humans cannot match the forte of the computer--its speed in examining large masses of data.

The solution to this problem is to let the computer do what it does well and let humans do what they do well. That is, design a mixed computer-human system, in which the computer handles the large mass of cases which can be classified as positive links or positive nonlinks, on a mechanical, routine basis. Carefully trained and well-motivated humans could then try to match the remaining cases, using a "computer-interactive" system, where the human would specify a small class of possible matches and the computer would display the records in this class, until a positive link was found or there was adequate evidence that no such link existed.

4. Techniques of Preprocessing

Certain elements of preprocessing will unquestionably be valuable in any computerized matching system. In particular, it is important to develop some method so that the computer can quickly and accurately identify the various elements of the name and address: surname, house number, street name or number, first name, and the conventional prefixes and suffixes to name and address. If this involves elaborate manual rearrangement and keying of the name and address, substantial error is likely to be introduced, possibly as much as the preprocessing removes. The examples in the paper suggest that unaided computer formatting is also likely to introduce as much error as it removes. A solution may be something used in one of the earliest (1956) computerized matching systems, where clerks inserted a distinctive and computer-readable symbol in front of the components of name and address to be used in the matching; e.g., * before surname, # before house number, % before street name, $ before P. O. box number, @ before title, etc. After appropriate codes were placed in fixed fields, the symbols were deleted from the computer records.

DISCUSSION

Benjamin J. Tepping, Westat, Inc.

The papers by Kirkendall and Kelley contain much interesting material, with some of which I must take issue.

The Fellegi-Sunter model, on which these papers are based, recognizes that there are three possible outcomes, but (it seems to me) uses the wrong utility function. To simply minimize the probability of subjecting a case to clerical review conditional on bounds on the probabilities of erroneous matches and erroneous nonmatches ignores important facts:

(a) the value of an erroneous match is, in many (or perhaps most) applications, quite different from the value of an erroneous nonmatch;

(b) the cost and the probability of misclassification associated with the clerical review should be taken into consideration.

We do not necessarily want to minimize the number of clerical reviews. We do want to maximize the value of the record linkage operation. This implies that one must not only determine the costs of the various components of the operation, but must also set values on the possible outcomes. An illustration of this approach is the application of a theoretical model of record linkage to the Chandrasekar-Deming technique for estimating the number of vital events on the basis of data from two different sources. This was published in the Bureau of the Census Technical Notes No. 4, in 1971 [1].

It appears that neither author is aware of my paper [2] in JASA in 1968 in which is presented a model for the optimum linkage of records.

The authors treat the problem as an exercise in the testing of hypotheses. I think it is preferable to regard it as a problem of decision making, subject to a utility function which depends upon the state of nature. In these applications, the three possible decisions are to call the pair of records being compared a match or a nonmatch, or to make some kind of further investigation before deciding on a classification. That investigation may consist simply of subjecting the records to personal scrutiny or may involve seeking additional data. The utility function would specify a gain or loss for each of the possible decisions, conditional on whether the pair is in fact a match or a nonmatch.

Kirkendall's examples also ignore the problem of fixing the values of the probabilities of errors of the first and second kinds. Those probabilities should not be arbitrary. Any solution of the problem should depend upon evaluation of the loss or gain of alternative decisions as well as on the cost of non-decisions--e.g., resort to other means of arriving at a decision.

Kirkendall's first illustration assumes independence, both under H_0 and under H_1. In the real world, this assumption may be far from true. For example, under either of the hypotheses H_0 or H_1, an agreement on first name would increase the probability of an agreement on the item sex--two records both giving the first name as "Nancy" are not likely to indicate different sexes. Presumably the lack of independence could be treated as in her example of cancer patients, essentially by dividing the First Name item into two items: one for cases in which both records show the sex as male and one for cases in which both records show the sex as female. This comment also applies to Kelley's numerical example, in which independence of these components is assumed.

As is pointed out by Kelley, the literature that gives advice on the choice of blocking schemes is not extensive. Yet practical problems make blocking of the files being compared essential, and Kelley's work should contribute to the improvement of blocking designs. He does take account of costs, by considering both the decrease in operational costs, because blocking reduces the number of comparison pairs, and the increase in the probability of an erroneous nonmatch as a result of blocking. (I note, however, that he does not use the fact that the probability of an erroneous match decreases as a result of the blocking.) His numerical examples illustrate that the choice among competing admissible blocking schemes involves the implicit assignment of relative values to an increase in the probability of erroneous nonmatches and a decrease in the number of comparisons. In practice, no doubt, a similar implicit assignment of values to an erroneous match, an erroneous nonmatch and a case referred to personal review is made in order to fix the values of the parameters λ and μ of the Fellegi- Sunter model.

I think there is difficulty with the application of Kelley's Lemma 2 to the determination of a suitable blocking scheme even after dealing with the lack of independence of the components of the comparison vector. It seems that a choice must depend, among other things, on a knowledge of the probability, given that the pair is a match (or a nonmatch), that there is agreement between the units of the pair on specified components of the comparison vector. Estimates of such probabilities must ultimately depend upon extensive empirical investigations, although such estimates seem often to be made on the basis of assumed models.

REFEPENCES

[1] Tepping, B.J., "The application of a linkage model to the Chandrasekar-Deming technique for estimating vital events," U.S. Bureau of the Census, Technical Notes No. 4, Washington, D.C., 1971, pp. 11-16.

[2] Tepping, B. J., "A model for optimum linkage of records," Journal of the American Statistical Association, 63, 1968, pp. 1321-1332.

REJOINDER

William E. Winkler, Energy Information Administration

Eli Marks' comments provide a valuable perspective to the overall objectives of matching procedures.

Just as the Fellegi-Sunter matching procedure contains computerized (automatic designation of matches and nonmatches) and manual (review of records designated for further manual followup) components, so does preprocessing contain computerized (minor reformatting, spelling standardization, string comparison) and manual (keypunch/transcription, major reformatting) components.

The respective roles of the two components are best exemplified by Newcombe et al. (1983, 1959, 1962). Newcombe's view is that computer procedures should be developed for the most routine and repetitive tasks. As knowledge of the characteristics of address files and coding techniques increases, computerized procedures can replace greater proportions -- possibly all -- manual components.

It is my experience that reasonably designed manual procedures are difficult and expensive to implement. This is because of high turnover rates and the necessity of training and constantly supervising personnel performing manual processing. Computerized procedures can have the benefit of being more cost-effective, consistent, and reproducible.

Both Marks and I note that the Census Bureau's ZIPSTAN software -- which is designed for files of individuals -- induced minor errors in files of businesses. In Winkler (1985), I show that ZIPSTAN's identification of address subfields can yield substantial improvements in the discriminating power of the Fellegi-Sunter matching procedure.

The cost in using ZIPSTAN was a few days of my time installing it. The alternative would have been to do nothing or develop manual procedures, set up computer files suitable for manual review, train individuals in computer login and manual review procedures, and have the individuals perform the review. Marks notes, if identifying individual subfields of the name and address involves "elaborate manual rearrangement and keying ..., substantial error is likely to be introduced, possibly as much as preprocessing removes."

I strongly agree that our understanding of "matching tolerances" needs to be improved. The purpose of my discussion of string comparators was to show the limitations of tolerances such as SOUNDEX, particularly SOUNDEX abbreviations of surnames used as sort keys during the blocking stage of matching. For files of businesses, I show (Winkler, 1985) that individual sort keys are generally not suitable for creating blocks containing most matched pairs. My solution is to apply independently multiple sort keys.

String comparison metrics, such as Jaro's string comparator, can only be efficiently used during the discrimination stage because they involve the comparison of corresponding strings from pairs of records. In my view, they offer the best opportunity for developing tolerances. How such tolerances fit in the framework of the Fellegi-Sunter model needs to be described and quantified.

REFERENCES

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959), "Automatic Linkage of Vital Records," Science 130, 954-959.

Newcombe, H.B. and Kennedy, J.M. (1962), "Record Linkage," Communications of the ACM 5, 563-566.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J., Strugnell, A., and Abbatt, J.D. (1983), "Reliability of Computerized Versus Manual Searches in a Study of the Health of Eldorado Uranium Workers," Comput. Biol. Med. 13, 157-169.

Winkler, W. E. (1985), "Exact Matching Lists of Businesses: Blocking, Subfield Identification, and Information Theory," Paper presented at the 1985 ASA Annual Meeting, Section on Survey Research Methods, August 4-8, 1985, Las Vegas, Nevada, (pp. 227-241 in this volume).

REJOINDER

R. Patrick Kelley, U.S. Bureau of the Census

Let me start my rejoinder by saying that I find Dr. Tepping's comments both interesting and helpful. The main criticism of my paper given by Dr. Tepping is my choice of the Fellegi-Sunter model as a basis for blocking research. As such, this exchange is simply another in a long debate over the handling of clerical costs and errors.

I have been aware of, and admired, Dr. Tepping's work on record linkage for quite some time. From a theoretical point of view, the utility theory approach is a fascinating one; however, clerical operations are hard to control and empirical investigations of clerical error rates and costs are data dependent. This makes estimates of the parameters in Dr. Tepping's model hard/expensive to obtain and highly variable.

Due to these facts, it is my opinion that the Fellegi-Sunter model provides the best general foundation for record linkage research and development. Methods which account for clerical costs should be used only after there have been several linkage projects run on data from the same source, using the same record linkage system.

Dr. Tepping also commented on the assumption of independence between comparison vector components, the difficulty of estimating, the difficulty of estimating model parameters, and the potential sensitivity of linkage error rates to errors in those parameter estimates. These comments are well placed, and I am continuing work on the blocking problem in an attempt to strengthen the results of this paper.

PROPERTIES OF THE SOCIAL SECURITY NUMBER RELEVANT TO ITS USE IN RECORD LINKAGES

Thomas B. Jabine, Consultant, Committee on National Statistics

Linkage of records from two data systems is aided greatly by the presence in both systems of the same numeric identifier, for example, the social security number (SSN) for persons or the employer identification number (EIN) for businesses. When matching variables for two records are compared, agreement on such numeric identifiers is usually given a large weight in deciding whether a true match exists.

Because of their importance for record linkage, it is important to have complete and current information on the relevant properties of each of these numeric identifiers. Such properties include: coverage, general structure and method of issuance, information content, and appropriate methods of validation. Properties relevant to sample selection using numeric identifiers are also of interest, since many record-linkage studies are based on a sample from one of the data systems.

This paper provides a description of the properties of the social security number (SSN) that are relevant to its use in record linkages. The description should be regarded as a first draft and readers are urged to suggest corrections and additions.

If this description of the SSN proves useful, it is suggested that the Administrative Records Subcommittee of the Federal Committee on Statistical Methodology make arrangements to: (1) prepare and disseminate descriptions, using the same format, of other commonly used numeric identifiers, such as the EIN and the unemployment insurance number, and (2) update the descriptions periodically and whenever significant changes occur.

Special thanks are due to Richard Wehrly of the Social Security Administration for providing information used in developing the SSN description. However, any errors are the sole responsibility of the author and readers are cautioned that the description of the SSN has not been officially reviewed by the Social Security Administration.

NUMERIC IDENTIFIER DESCRIPTION

1. <u>Name of identifier</u>

<u>The social security</u> number (SSN).

2. <u>Administrative uses</u>

<u>SSNs were issued i</u>nitially so that earnings of persons in jobs covered by the social security retirement program could be reported, by their employers, to the Social Security Administration (SSA) and credited to the persons accounts for subsequent use in determining benefit eligibility and payment amounts.

An early decision was made to use SSNs as identifiers in the State-operated unemployment insurance programs. No other significant uses developed until 1961 when the Internal Revenue Service, after discussions with SSA, decided to use the SSN as a taxpayer identification number. After implementation of this decision, other uses by Federal and State governments followed rapidly, and the SSN is now widely used as an identifier for workers, taxpayers, drivers, students, welfare beneficiaries, civil servants, servicemen, veterans, pensioners and others (HEW Secretary's Advisory Committee, 1973).

Legal justification for use of the SSN as an identifier by Federal agencies comes from Executive Order 9397, issued in 1943, which directed Federal agencies to use the SSN when establishing a new system of permanent account numbers. The Privacy Act of 1974 placed some restrictions on use of SSNs by Federal, State and local government agencies, but uses formally established prior to January 1, 1975 were not affected and these restrictions have had only a minor effect on widespread administrative use of the SSN by governments and private organizations (Privacy Protection Study Commission, 1977).

3. <u>Coverage</u>

a. <u>Units</u>.--SSNs are issued to persons.

b. <u>Legal coverage provisions</u>.--An SSN will be issued to any United States citizen upon application and presentation of acceptable evidence of identity. Foreign nationals legally present in the United States will be issued SSNs if legally entitled to work or if they have an acceptable "nonwork reason" for needing an SSN, e.g., the need for a taxpayer identification number.

All persons with Federally taxable income and their spouses are required to obtain SSNs for use as taxpayer identification numbers. SSNs are also required for many types of benefits and for other purposes: social security, driver's license, welfare benefits, voter registration, participation in scholastic aptitude testing programs, etc. For some of these, requirements vary by State.

c. <u>Volume and characteristics of issuance to date</u>.--SSNs were first issued in November 1936. By the end of 1975, over 235 million SSNs had been issued and there were an estimated 180 million living SSN holders (Social Security Administration, 1981b). As of the close of 1983, approximately 287,083,000 SSNs had been issued. It is estimated by SSA that there were 204,760,000 living SSN holders at the end of 1981. When SSN holders die, their SSNs are <u>not</u> reissued to other applicants.

The table in Attachment A shows the number of SSNs issued annually, by sex of applicant, through the end of 1979. Following the large number of issuances in the first 14 months (November 1936 to December 1937), the volume of annual issuances has fluctuated for a variety of reasons, with a tendency to increase in recent years as coverage of SSA benefit programs and the use of SSNs for non-SSA programs has expanded. Today most of the SSNs are issued to applicants under 20 years of age. In 1979, 62.8 percent of the SSNs were issued to persons under 15 and another 26.2 percent to persons between 15 and 19 (Social Security Administration, 1981b).

From time to time, surname counts based on the first six characters of the surname are made from SSA's account number files. Kilss and Tyler (1974) show the rankings of common surnames based on 1964 counts. Based on a 1974 tabulation, the ten most common surnames were:

Smith
Johnso(n)
Willia(ms)(mson)
Brown
Jones
Miller
Davis
Martin(ez)(son)
Anders(on)
Wilson

The letters in parentheses following some names are intended to show the more common surnames that have these first six characters.

d. Uniqueness, stability.--Until 1972, applicants for SSNs were not asked if they had already been issued numbers, nor were they asked for proof of identity. As a result many persons now have more than one SSN (Privacy Protection Study Commission, 1977). As of 1973, it was estimated that 4.2 million persons had two or more SSNs (HEW Secretary's Advisory Committee, 1973). More recent estimates are not available. Today, intentional issuance of multiple numbers to the same person is permitted only in exceptional circumstances, generally involving national security or the protection of the person in question.

In most cases where a person is known to have more than one SSN, SSA's computerized SSN files contain a record for each of his or her SSNs and cross references linking all of the SSNs.

Sometimes more than one person uses the same SSN. Some reasons why this happens are discussed in item 8b. Estimates of the frequency with which this occurs are not readily available, but it is believed to be much less prevalent than issuance of multiple numbers to the same person (HEW Secretary's Advisory Committee, 1973).

4. General structure and information content

The social security number has nine digits arranged as follows: 000-00-0000. The first three digits are called the area number, the next two are the group number, and the last four are the serial number. There are no check digits. The serial number provides no information about the person to whom an SSN has been assigned; however, the area and group numbers do contain a limited amount of information.

The area number, digits one to three of the SSN, carries some information either about the SSN holder's occupation or his or her place of residence at the time the number was issued. For the ranges of area numbers used to date, the information content is as follows:

(1) Area numbers 001 to 626. With a few exceptions, each of these area numbers has been assigned to a single State, one or more to a State. For most SSNs, the area number indicates only the SSN holder's State of residence at the time of issuance, as derived from the mailing address on the SSN application. For SSNs issued in the early days of social security, the area number indicated the specific SSA field office from which the number was issued, regardless of where the applicant lived.

(2) Area numbers 700-728. These numbers were assigned to railroad workers through 1963. Since then, railroad workers have been assigned SSNs with the same area numbers as other applicants.

The group number, digits four and five, in combination with the area number, provides a rough indication of when the SSN was issued. In particular, it is possible to tell whether an SSN was issued before or after another SSN having the same area number but a different group. Within an area number, the group numbers are always used in the following sequence:

- Odd numbers from 01 to 09
- Even numbers from 10 to 98
- Even numbers from 02 to 08
- Odd numbers from 11 to 99

The group number 00 has never been used. Only the first two sets of group numbers in the above sequence were used through 1965. Since then the third and fourth sets have been used with some area numbers. Current information on the last group number assigned for each area number can be obtained from SSA (see Section 9.a.).

5. Issuance procedures

All SSNs are issued by the Social Security Administration. Prior to July 1, 1963, the Railroad Retirement Board issued SSNs (in the 700 series) to all railroad employees.

A single application form, Form SS-5, Application for a Social Security Number Card, is used for initial applications, requests for replacements for lost cards and corrections, such as name changes. A copy of the application form is shown in Attachment B. Applications must be accompanied by evidence of age, identity and U.S. citizenship or lawful alien status. They may be submitted either in person or by mail, except that aliens and persons 18 or older making initial applications must apply in person.

Most SSN applications are submitted to SSA field offices. In 37 States, applications for new welfare applicants needing SSNs are developed by the State welfare agencies and submitted by the State directly to SSA's Office of Central Records Operations. SSA district offices sometimes make arrangements with schools for "mass enumerations" in which SSA and school officials collaborate in obtaining and reviewing applications from all students who wish to obtain SSNs.

The application forms (SS-5) and accompanying evidence submitted to district offices are screened for completeness and accuracy by district office personnel, who make further contacts with applicants when necessary. The SS-5 information is then keyed in the district office for direct transmission to SSA central operations.

The central processing of the applications consists of validation (which is essentially a matching operation) against existing SSN files, followed by appropriate actions. The exact nature of the validation depends on the type of application. For example, if an initial applicant alleges that he or she has not been issued an SSN previously, the purpose of the validation is to confirm that allegation. Validation procedures are discussed further in item 9b.

The final step depends on the results of the validation. The main possibilities are: assigning an SSN and mailing a card to a new applicant, mailing a replacement card to an applicant, correcting information (such as name) about the applicant in the SSN computerized files, or asking the field office to supply additional information.

When a new SSN is assigned, the next available number for the State from which the application was submitted is used. The sequence of availability proceeds from the lowest area number used in a given State through the highest area number for that State, using the same group number. For example, in New Hampshire, which has been assigned area codes 001, 002, and 003, the last available number in group 001-52 would be followed by the first available number in group 002-52, and the last available number in that group would be followed by the first available number in group 003-52.

6. Sampling properties

In theory, a probability sample could be selected using digital patterns based on any of the nine digits of the SSN or combinations thereof. However, consideration of the information content of the first five digits, as described in item 4, makes it clear that use of any of those digits should be avoided. It would be most inconvenient to select a sample that turned out to include only persons who were railroad workers at the time their SSNs were issued and had all been issued their SSNs not later than 1963!

The serial number part of the SSN, however, does not have this kind of problem and consequently is frequently used for digital sampling from a file of records that includes SSNs. Assuming a uniform distribution of 9,999 possible serial numbers (SSNs ending in 0000 have never been issued), it is possible to choose a digital sampling pattern that will approximate any desired sampling fraction. There are usually several alternatives. For example, to select a sample of approximately 5 percent (1 in 20) of the records, one could use

(1) 5 of the 100 possible combinations of the 8th and 9th digits;

(2) 50 of the 1,000 possible combinations of digits 7, 8 and 9;

(3) 500 of the 9,999 combinations of digits 6, 7, 8 and 9;

(4) 5 of the 100 possible combinations of the 7th and 8th digits

and so forth. The combinations of digits selected may be chosen at random with or without replacement (the latter would be preferable) or systematically with a random start. In the latter case, for exmple, we might choose the pair 73 at random and include with it the pairs 93, 13, 33 and 53.

The use of selected digits or combinations of digits for sampling is actually a form of cluster sampling. In the illustration used above, we could describe a population of records as consisting of 100 clusters, each consisting of all records with SSNs having a particular pair of 8th and 9th digits. Five of these clusters are selected by an appropriate probability sampling mechanism.

In practice, samples of this kind, especially when only the 8th and 9th digits are used, behave pretty much like random samples, chosen without replacement. In particular, reasonably accurate estimates of sampling error can be calculated as though the data were from a simple random sample.

In selecting samples based on the serial number portion of the SSN, the following points should be considered:

(1) The serial number 0000 is not used. The effect of this, which is quite small, on the expected sample size can easily be calculated.

(2) The digital patterns used for any particular sample determine only the expected sampling fraction or size. The sample size realized by using a particular set of digits or combination of digits will, in general, differ somewhat from its expected value. If precise control of sample size is important, this can be achieved by oversampling initially and then subsampling units at random or systematically from the initial sample.

(3) As discussed in item 3d, some persons have been issued more than one SSN. Such persons may have multiple chances of selection in a sample of persons obtained by selecting SSNs, depending on what record sets are being used. If the number of SSNs that each sample person has can be determined, appropriate adjustments can be made in estimates based on the sample. Because the phenomenon is infrequent, however, it is usually ignored in practice.

(4) Various studies (Hawkes and Harris, 1969; Page and Wright, 1979) have shown that the distributions of SSNs by ending digit in selected record sets is essentially uniform. However, studies conducted with various record sets in the late 1960s and early 1970s (Hawkes and Harris, 1969; Internal Revenue Service, 1973) showed a negative linear relationship between the ascending sequence of digits in positions 6 and 7 and the number of SSNs in these record sets having those digits. This probably resulted from the fact that, until 1972, SSNs in each area-group combination were issued consecutively by serial number, from 0001 to 9999. Since then, they have been issued in a randomized order, largely to avoid issuing consecutive numbers to persons with the same surname. Because of the new issuance procedure, one would expect this relationship to disappear gradually. However, to be on the safe side, it is recommended that: (1) digital sampling patterns use only the 8th and 9th digits whenever requirements can be met in that way, and (2) whenever multiple combinations of two or more digits are used, they should be selected systematically rather than at random from the range of possible combinations.

7. Links with other numeric identifiers

At the Federal level, there are two kinds of links between SSNs and employer identification numbers (EINs). For employees, the link occurs in the W-2/W-3 annual wage and tax reporting system (prior to 1978, reporting was quarterly). For many years SSA has used this link for statistical purposes, in the Continuous Work History Sample system, to add employer locations and industry data to records of earnings and demographic characteristics for sample persons. More recently, the Statistics of Income Division of IRS has used the same link to obtain employer industry codes to use as an aid in coding occupations reported by individual taxpayers on their returns.

The second link between SSNs and EINs applies to persons who operate businesses as sole proprietors. This link applies primarily to sole proprietors with employees; those with no employees are not, in general, required to obtain and use EINs. The link occurs in two ways: on income tax returns of sole proprietors, and on new applications for EINs. On income tax returns, the business schedules (C and F) call for entries of both the EIN (if the taxpayer has one) and the SSN. On EIN application forms (Form SS-4), applicants who are sole proprietors are asked to enter their SSNs.

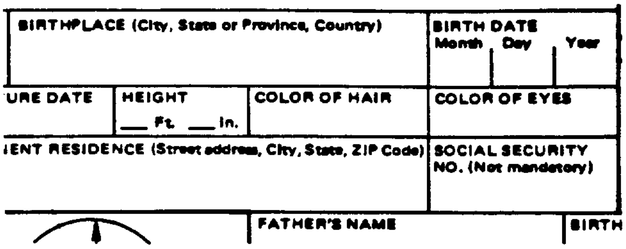
There are undoubtedly several links between the SSN and other numeric identifiers at the State and local levels. One obvious one is the link between SSNs and employer unemployment insurance (UI) identification numbers, which is necessary for the operation of the UI program. The precise nature of the linkage varies by State and, for the minority of States which operate under the "wage request" system, it may not exist in any readily accessible sense.

8. Reporting formats and problems

a. Formats.--Many different administrative and statistical forms include spaces for recording SSNs, either by the holders or by someone else completing the form. There is no standard format for this purpose. The particular format used may have some effect on the accuracy with which SSNs are entered on the forms and read from the forms for purposes of manual transcription or data entry.

Format features that vary include: width and height of the space provided for the number; separators used for the area, group, and serial numbers; use of boxes for individual digits; and the label used to indicate what should be entered. Some examples of these features appear below. All of them show the actual size of the entry space on the form.

Example 1. Department of State, Passport Application, Form SDP-11 (7-79)



BIRTHPLACE (City, State or Province, Country)

BIRTH DATE
Month | Day | Year

URE DATE

HEIGHT
___ Ft. ___ In.

COLOR OF HAIR

COLOR OF EYES

IENT RESIDENCE (Street address, City, State, ZIP Code)

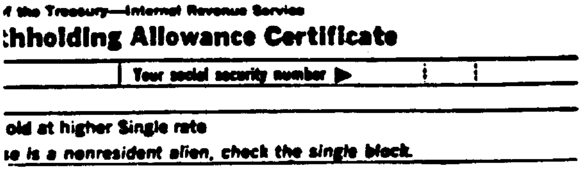
SOCIAL SECURITY NO. (Not mandatory)

FATHER'S NAME

BIRTH

Of several formats examined, this one provided the narrowest space for entering the SSN, with a width of 1 1/4 inches. Most others were in the range of 1 1/2 to 2 inches.

Example 2. Internal Revenue Service, Employee's Withholding Allowance Certificate, Form W-4 (10-79)



f the Treasury—Internal Revenue Service

thholding Allowance Certificate

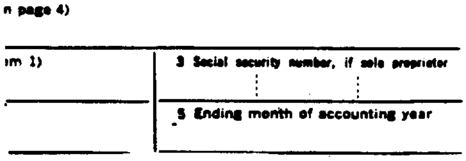
Your social security number ▶ ⋮ ⋮

old at higher Single rate

se is a nonresident alien, check the single block.

This format allowed the smallest vertical distance of those examined, 5/32 inch. It uses vertical dotted lines as separators for the three parts of the SSN.

Example 3. Internal Revenue Service, Application for Employer Identification Number, Form SS-4, (8-76).



n page 4)

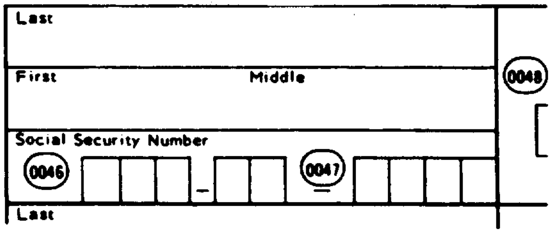
rm 1)

3 Social security number, if sole proprietor

5 Ending month of accounting year

This format also uses the dotted vertical lines as separators. In this case, the spaces for the three portions of the SSN are all the same length, 5/8 inch. Other forms using separators make the lengths of the three spaces roughly proportional to the number of digits to be entered, i.e., 3, 2, and 4.

Example 4. Bureau of the Census/Department of Health and Human Services, Income Survey Development Program, 1978 Research Panel-July Questionnaire, Form ISDP-403.



Last

First Middle

0048

Social Security Number

0046 □□□-□□ 0047 -□□□□

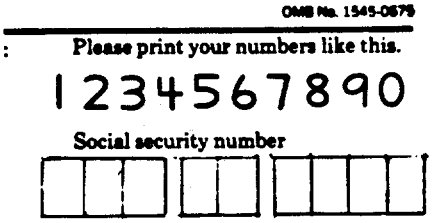
Last

This format illustrates the use of separate boxes for each digit of the SSN. The three parts of the SSN are separated by horizontal dashes. The circled numbers are source codes for data entry.

Example 5. Social Security Number Card (Original, Replacement or Correction), Form SS-5 (5-84) (see Attachment B).

This item is completed only for persons who already have SSNs and are applying for a replacement or correction. This format uses a box for each digit, with intervening spaces, and horizontal dashes to separate the three parts of the SSN. The wording of the item label reflects the fact that the form is sometimes completed by someone other than the "applicant."

Example 6. Internal Revenue Service, Form 1040 EZ Income Tax Return for Single Filers with no Dependents.



OMB No. 1545-0675

Please print your numbers like this.

1234567890

Social security number

□□□ □□ □□□□

This format is used for handwritten entries by taxpayers that will be read automatically by optical character reading equipment. On the actual form, the boxes for the individual digits are in light blue. The boxes for the area, group and serial parts of the SSN are separated.

Example 4 above comes from a questionnaire that is completed by trained Census Bureau interviewers. The other examples are all from forms that are filled by members of the general public. No experimental research on alternative formats for recording SSNs has been identified. Some other research has suggested that the use of individual character separators may actually reduce legibility of entries (Wright, 1980).

b. Reporting and processing errors.--Most errors in SSNs in data files occur for two reasons: (1) the person completing the form or answering the questions gave an SSN for the wrong person, or (2) the SSN is for the right person, but it was reported, recorded, transcribed or keyed incorrectly.

The first type of error can occur, for example, when a widow reports the number under which she is receiving benefits, rather than her own. Another example is what SSA calls the "pocketbook number." The number 078-05-1120 appeared on a sample account number card contained in wallets sold nationwide in 1938. Several thousand people mistakenly reported this number to their employers as their own! By the 1970s there were over 20 different pocketbook numbers (HEW Secretary's Advisory Committee, 1973, p. 112).

People who lose their social security cards can apply for replacement cards bearing the SSN already issued to them. In cases where they are not able to give their SSN on the application, SSA must determine the correct SSN based on other identifying information. Occasionally a mismatch occurs and the person will be issued a replacement card bearing someone else's SSN.

The second type of error is usually an error in a single digit or a transposition of digits, types of errors that could be easily corrected if a check digit were used.

Cobleigh and Alvey (1974) describe errors detected when SSNs reported in the Current Population Survey were validated against Social Security Administration files. About three percent of the reported SSNs were clearly in error. Roughly two-thirds of these were found to have transposition or single-digit errors. Another one-sixth were SSNs belonging to other members of the same household, and the remainder could not be located in SSA's files.

9. Validation procedures

a. Intra-record validation.--When undertaking record linkages based on SSNs, it is usually desirable to start by identifying SSNs that are clearly invalid. A first step might be to look at the SSN itself and determine whether it is within the range of numbers issued to date. SSA will make available, on request, up-do-date information on the area numbers that have been issued so far and, for each of those numbers, the "highest" group number issued. "Highest" must be interpreted in terms of the standard sequence for use of group numbers within an area number, as explained in item 4 above.

Attachment C provides this information as of January 2, 1985. As of that date, the only area numbers used were those in the ranges 001 to 587, 589 to 595, 600 and 601, and 700 to 728. Also, group number 00 and serial number 0000 are never used. Current information on highest group numbers may be obtained from the director of the OASDI Statistics Division; Office of Research, Statistics and International Policy; Social Security Administration.

If records to be linked have information on date of birth or age, the SSN can be checked for consistency with age. The operating rule is that a person whose SSN was issued x years ago must be at least x years old. Since virtually all numbers issued through 1961 were issued to employed persons, only a few errors would be made by requiring that persons with numbers issued in this period be at least x + 15 years old. For SSNs issued from 1951 onwards, the SSA can provide fairly precise information about the years in which numbers with specific area-group combinations were issued (contact the source given in the preceding paragraph). For numbers issued prior to 1951, only rough estimates of issuance periods for area-group combinations are possible.

b. Validation against SSA records.--Validation is defined broadly here as a process in which SSN information for individuals from sources external to SSA records is checked against those records to determine its validity. Specifically, if the external record includes an SSN, it is desired to know whether the SSN is the correct one for that person and, if it is not correct, what the correct SSN, if any, is for that person. If the external record for a person has no SSN, it is desired to know whether that person has an SSN and, if so, what it is. This kind of validation requires matching external records to SSA records and should be thought of in that context.

Validation of SSN information is done routinely by SSA for program purposes. Somewhat less frequently it is undertaken for statistical purposes. Some examples of the latter are:

(1) Validation of SSNs collected in pretests for the 1970 Census of Population (Ono et al., 1968).

(2) Validation of SSNs collected in the March 1973 Current Population Survey, as a preparatory step before adding SSA and IRS administrative data to the survey records (covered in several reports and articles, e.g., Cobleigh and Alvey, 1974; Social Security Administration, 1981a).

(3) Validation of SSNs collected in panel surveys as part of the Income Survey Development Program (Kasprzyk, 1983).

(4) In various mortality followup studies, as a preparatory step before determining which members of an externally identified study population have died, according to SSA records.

Attachment D provides a summary description of SSA's current validation procedures for program operations. A combination of computerized and manual procedures is used, and unresolved cases are returned to district offices with an instruction to seek additional information from the applicant or claimant. The SSN files maintained by SSA are now fully computerized and a more sophisticated computer validation system is being developed.

A variety of validation procedures have been used in statistical applications; some of them are described in the references cited above.

The circumstances under which SSA will validate SSN information for administrative or statistical purposes are limited by law and by SSA regulations and policies. Anyone wishing to validate SSN information for statistical or research purposes should contact SSA's Office of Research, Statistics and International Policy.

10. Use as a matching variable

Arellano (n.d.) discusses use of the SSN in record linkages based on the model proposed by Fellegi and Sunter (1969). He recommends that the SSN not be used for blocking, because of the possibility that some individuals in the files to be linked may not have been issued SSNs. To use the SSN as a component of the comparison vector, Arellano recommends that the 9 digits of the SSN be partitioned into four elements on a 2,2,2,3 basis. He identifies 17 possible configurations of the SSN component of the comparison vector, covering the possible realizations of agreements and disagreements in the four elements, plus the case in which no SSN is available for one or both members of the comparison pair. He then suggests procedures for assigning conditional probabilities to these configurations for the matched and unmatched sets. These probabilities are based on assumptions about the kinds of errors that can occur in the matched set and on observed frequencies of realizations of the first three elements of the partitioned SSNs in the files to be linked (realizations of the fourth element are assumed to be uniformly distributed).

Rogot et al. (1983) report on linkages of records from the Census Bureau's Current Population Survey with the National Death Index, using each person's name, SSN and date of birth as key matching variables. Based on the results of an evaluation study in which "truth" (match or non-match) was based on a consensus of three raters using all available information for a set of "possible matches," they concluded that whenever SSNs agreed, it was appropriate to classify the pair of records as a positive link, provided there was agreement on sex. The use of probabilistic matching procedures was restricted to cases for which the SSNs did not agree or were missing on one or both records.

REFERENCES

Arellano, M.
(n.d.) Optimum utilization of the social security number for matching purposes. No further identification available.

Cobleigh, C. and Alvey, W.
1974 Validating reported social security numbers. American Statistical Association Proceedings, Social Statistics Section, 145-150.

Fellegi, I. and Sunter, A.
1969 A theory for record linkage. Journal of the American Statistical Association, 64(328); 1183-1210.

Hawkes, T. and Harris, R.
1969 An analysis of social security numbers in the SMI actuarial sample. Actuarial Note No. 62. Social Security Administration, U.S. Department of Health, Education, and Welfare.

HEW Secretary's Advisory Committee on Automated Personal Data Systems
1973 Records, Computers and the Rights of Citizens. Washington, D.C.: U.S. Department of Health, Education, and Welfare.
Chapt. VII. The Social Security Number as a Standard Universal Identifier
Chapt. VIII. Recommendations Regarding Use of the Social Security Number

Internal Revenue Service
1973 Evaluation of the randomness of the four ending digits of the social security numbers on the master file, 1971. Unpublished report by Mathematical Statistics Branch, Statistics Division. U.S. Department of the Treasury.

Kasprzyk, D.
1983 Social security number reporting, the use of administrative records, and the multiple frame design in the Income Survey Development Program. Pp. 123-144 in Technical, Conceptual and Administrative Lessons of the ISDP (M. David, ed.) Washington, D.C.: Social Science Research Council.

Kilss, B. and Tyler, B.
1974 Searching for missing social security numbers. American Statistical Association Proceedings, Social Statistics Section, 137-144.

Ono, M., Patterson, G. and Weitzman, M.
1968 The quality of reporting social security numbers in two surveys. American Statistical Association Proceedings, Social Statistics Section, 197-205.

Page, W. and Wright, G.
1979 A Statistical Study of the VA Annual Patient Census Sampling Procedure, 1975-1977, Controller Monograph Technical Series, No. 1. Veterans Administration.

Privacy Protection Study Commission
1977 Personal Privacy in an Information Society. Washington, D.C.: U.S. Government Printing Office.
Chapt. 16. The Social Security Number.

Rogot, E., Schwartz, S., O'Conor, K. & Olsen, C.
1983 The use of probabilistic methods in matching census samples to the National Death Index. Pp. 75-80 in Statistics of Income and Related Administrative Record Research: 1983, Internal Revenue Service.

Social Security Administration
1981a Methods of Estimation for the 1973 Exact Match Study. Studies from interagency data linkages, Report No. 10. Washington, D.C.: Department of Health and Human Services.

Social Security Administration
1981b Social security numbers issued, 1937-79. Research and Statistics Notes, No. 7, by F. Bamberger. Washington, D.C.: U.S. Department of Health and Human Services.

Wright, P.
1980 Strategy and tactics in the design of forms. Visible Language 14(2): 151-193.

ATTACHMENT A

Table 1.--Social Security Numbers Issued, By Sex of Applicants, 1937-79

(In thousands)

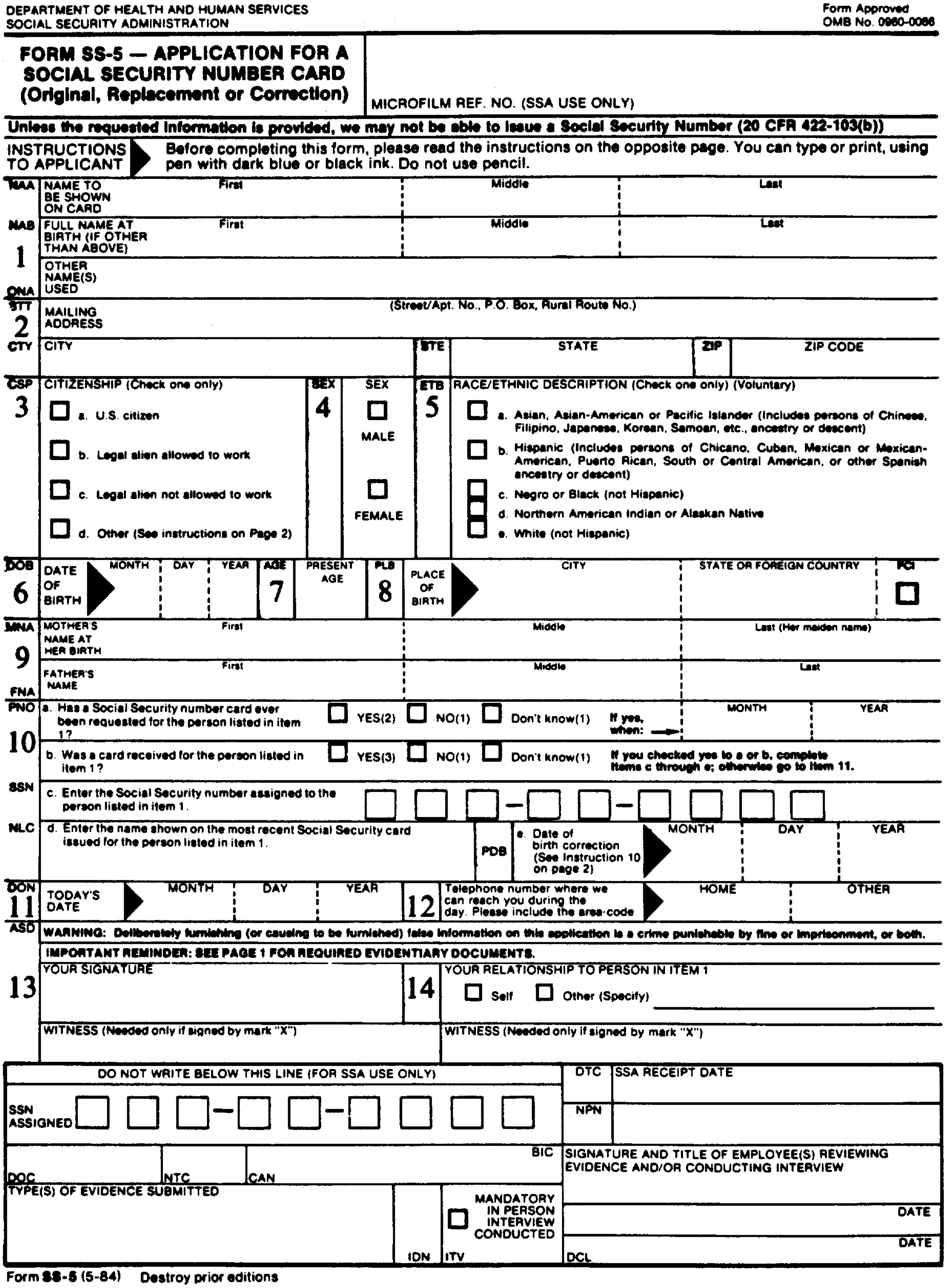
Year	Total	Male	Female
1937[1]	37,139	26,981	10,158
1938	6,304	4,010	2,294
1939	5,555	3,291	2,264
1940	5,227	3,080	2,147
1941	6,678	3,702	2,976
1942	7,637	3,547	4,090
1943	7,426	2,905	4,521
1944	4,537	1,830	2,707
1945	3,321	1,506	1,815
1946	3,022	1,432	1,590
1947	2,728	1,299	1,429
1948	2,720	1,305	1,415
1949	2,340	1,113	1,227
1950	2,891	1,406	1,485
1951	4,927	2,420	2,507
1952	4,363	2,292	2,071
1953	3,464	1,664	1,800
1954	2,743	1,299	1,444
1955	4,323	2,304	2,019
1956	4,376	2,391	1,985
1957	3,639	1,793	1,846
1958	2,920	1,384	1,536
1959	3,388	1,645	1,743
1960	3,415	1,663	1,752
1961	3,370	1,665	1,705
1962	4,519	2,109	2,410
1963	8,617	3,739	4,878
1964	5,623	2,707	2,916
1965	6,131	2,746	3,385
1966	6,506	2,894	3,612
1967	5,920	2,855	3,065
1968	5,862	2,856	3,006
1969	6,289	3,105	3,184
1970	6,132	3,004	3,128
1971	6,401	3,122	3,279
1972	9,564	3,948	5,616
1973	10,038	4,849	5,189
1974	7,998	3,950	4,048
1975	8,164	3,992	4,172
1976	9,043	4,507	4,536
1977	7,724	3,872	3,852
1978	5,260	2,682	2,578
1979	5,213	2,649	2,564

[1] Includes issuances in November and December 1936.

Source: Social Security Administration, 1981b.

ATTACHMENT B

Form SS-5.--Application for a Social Security Number Card



DEPARTMENT OF HEALTH AND HUMAN SERVICES
SOCIAL SECURITY ADMINISTRATION

Form Approved
OMB No. 0960-0066

FORM SS-5 — APPLICATION FOR A SOCIAL SECURITY NUMBER CARD (Original, Replacement or Correction)

MICROFILM REF. NO. (SSA USE ONLY)

Unless the requested information is provided, we may not be able to issue a Social Security Number (20 CFR 422-103(b))

INSTRUCTIONS TO APPLICANT ▶ Before completing this form, please read the instructions on the opposite page. You can type or print, using pen with dark blue or black ink. Do not use pencil.

1
NAA NAME TO BE SHOWN ON CARD — First | Middle | Last
NAB FULL NAME AT BIRTH (IF OTHER THAN ABOVE) — First | Middle | Last
ONA OTHER NAME(S) USED

2
STT MAILING ADDRESS (Street/Apt. No., P.O. Box, Rural Route No.)
CTY CITY | STE STATE | ZIP ZIP CODE

3 CSP CITIZENSHIP (Check one only)
□ a. U.S. citizen
□ b. Legal alien allowed to work
□ c. Legal alien not allowed to work
□ d. Other (See instructions on Page 2)

4 SEX SEX
□ MALE
□ FEMALE

5 ETB RACE/ETHNIC DESCRIPTION (Check one only) (Voluntary)
□ a. Asian, Asian-American or Pacific Islander (Includes persons of Chinese, Filipino, Japanese, Korean, Samoan, etc., ancestry or descent)
□ b. Hispanic (Includes persons of Chicano, Cuban, Mexican or Mexican-American, Puerto Rican, South or Central American, or other Spanish ancestry or descent)
□ c. Negro or Black (not Hispanic)
□ d. Northern American Indian or Alaskan Native
□ e. White (not Hispanic)

6 DOB DATE OF BIRTH ▶ MONTH | DAY | YEAR
7 AGE PRESENT AGE
8 PLB PLACE OF BIRTH ▶ CITY | STATE OR FOREIGN COUNTRY | FCI □

9
MNA MOTHER'S NAME AT HER BIRTH — First | Middle | Last (Her maiden name)
FNA FATHER'S NAME — First | Middle | Last

10
PNO a. Has a Social Security number card ever been requested for the person listed in item 1? □ YES(2) □ NO(1) □ Don't know(1) If yes, when: → MONTH | YEAR
b. Was a card received for the person listed in item 1? □ YES(3) □ NO(1) □ Don't know(1) **If you checked yes to a or b, complete items c through e; otherwise go to item 11.**
SSN c. Enter the Social Security number assigned to the person listed in item 1. □□□–□□–□□□□
NLC d. Enter the name shown on the most recent Social Security card issued for the person listed in item 1.
PDB e. Date of birth correction (See Instruction 10 on page 2) ▶ MONTH | DAY | YEAR

11 DON TODAY'S DATE ▶ MONTH | DAY | YEAR
12 Telephone number where we can reach you during the day. Please include the area code ▶ HOME | OTHER

ASD **WARNING: Deliberately furnishing (or causing to be furnished) false information on this application is a crime punishable by fine or imprisonment, or both.**

IMPORTANT REMINDER: SEE PAGE 1 FOR REQUIRED EVIDENTIARY DOCUMENTS.

13 YOUR SIGNATURE
14 YOUR RELATIONSHIP TO PERSON IN ITEM 1 □ Self □ Other (Specify) ____________

WITNESS (Needed only if signed by mark "X")
WITNESS (Needed only if signed by mark "X")

DO NOT WRITE BELOW THIS LINE (FOR SSA USE ONLY)

SSN ASSIGNED □□□–□□–□□□□

DTC SSA RECEIPT DATE
NPN

DOC | NTC | CAN | BIC

SIGNATURE AND TITLE OF EMPLOYEE(S) REVIEWING EVIDENCE AND/OR CONDUCTING INTERVIEW
DATE
DATE
DCL

TYPE(S) OF EVIDENCE SUBMITTED
IDN
□ MANDATORY IN PERSON INTERVIEW CONDUCTED
ITV

Form SS-5 (5-84) Destroy prior editions

ATTACHMENT C

Distribution of Social Security Numbers as of January 2, 1985: Highest Group Number Issued Within Each Area Number*

		059	68	118	68	177	64	236	25	295	80	354	72	413	45
001	68	060	68	119	68	178	64	237	45	296	80	355	72	414	45
002	68	061	68	120	68	179	64	238	45	297	80	356	72	415	43
003	66	062	68	121	68	180	64	239	45	298	80	357	72	416	19
004	82	063	68	122	66	181	64	240	43	299	80	358	72	417	19
005	80	064	68	123	66	182	64	241	43	300	80	359	72	418	19
006	80	065	68	124	66	183	64	242	43	301	80	360	72	419	19
007	80	066	68	125	66	184	64	243	43	302	80	361	72	420	19
008	66	067	68	126	66	185	64	244	43	303	92	362	94	421	19
009	64	068	68	127	66	186	64	245	43	304	92	363	94	422	19
010	66	069	68	128	66	187	64	246	43	305	92	364	94	423	19
011	66	070	68	129	66	188	64	247	59	306	92	365	94	424	17
012	64	071	68	130	66	189	64	248	59	307	92	366	94	425	51
013	64	072	68	131	66	190	64	249	59	308	92	367	94	426	51
014	64	073	68	132	66	191	64	250	57	309	92	368	94	427	49
015	64	074	68	133	66	192	64	251	57	310	92	369	94	428	49
016	64	075	68	134	66	193	64	252	49	311	92	370	94	429	57
017	64	076	68	135	78	194	64	253	49	312	92	371	94	430	57
018	64	077	68	136	78	195	64	254	49	313	92	372	94	431	55
019	64	078	68	137	78	196	64	255	49	314	92	373	94	432	55
020	64	079	68	138	76	197	64	256	49	315	92	374	94	433	55
021	64	080	68	139	76	198	64	257	47	316	92	375	94	434	55
022	64	081	68	140	76	199	64	258	47	317	92	376	94	435	55
023	64	082	68	141	76	200	62	259	47	318	74	377	94	436	55
024	64	083	68	142	76	201	62	260	47	319	74	378	94	437	55
025	64	084	68	143	76	202	62	261	99	320	74	379	94	438	55
026	64	085	68	144	76	203	62	262	99	321	74	380	94	439	53
027	64	086	68	145	76	204	62	263	99	322	74	381	94	440	84
028	64	087	68	146	76	205	62	264	99	323	74	382	94	441	84
029	64	088	68	147	76	206	62	265	99	324	74	383	92	442	84
030	64	089	68	148	76	207	62	266	99	325	74	384	92	443	84
031	64	090	68	149	76	208	62	267	99	326	74	385	92	444	84
032	64	091	68	150	76	209	62	268	82	327	74	386	92	445	84
033	64	092	68	151	76	210	62	269	82	328	74	387	92	446	82
034	64	093	68	152	76	211	62	270	82	329	74	388	92	447	82
035	54	094	68	153	76	212	06	271	82	330	74	389	92	448	82
036	52	095	68	154	76	213	06	272	82	331	74	390	92	449	69
037	52	096	68	155	76	214	06	273	82	332	74	391	92	450	69
038	52	097	68	156	76	215	06	274	82	333	74	392	92	451	69
039	52	098	68	157	76	216	06	275	82	334	74	393	92	452	69
040	76	099	68	158	76	217	06	276	82	335	74	394	92	453	69
041	76	100	68	159	64	218	06	277	82	336	74	395	92	454	69
042	76	101	68	160	64	219	06	278	82	337	74	396	92	455	69
043	76	102	68	161	64	220	04	279	82	338	74	397	92	456	69
044	76	103	68	162	64	221	68	280	82	339	74	398	92	457	69
045	76	104	68	163	64	222	66	281	82	340	74	399	92	458	69
046	76	105	68	164	64	223	33	282	82	341	74	400	25	459	69
047	76	106	68	165	64	224	33	283	82	342	72	401	25	460	69
048	76	107	68	166	64	225	33	284	82	343	72	402	25	461	69
049	74	108	68	167	64	226	33	285	82	344	72	403	25	462	69
050	68	109	68	168	64	227	33	286	82	345	72	404	25	463	69
051	68	110	68	169	64	228	33	287	82	346	72	405	25	464	69
052	68	111	68	170	64	229	33	288	82	347	72	406	23	465	69
053	68	112	68	171	64	230	31	289	82	348	72	407	23	466	69
054	68	113	68	172	64	231	31	290	80	349	72	408	45	467	69
055	68	114	68	173	64	232	27	291	80	350	72	409	45	468	04
056	68	115	68	174	64	233	27	292	80	351	72	410	45	469	04
057	68	116	68	175	64	234	27	293	80	352	72	411	45	470	04
058	68	117	68	176	64	235	25	294	80	353	72	412	45	471	04

Distribution of Social Security Numbers as of January 2, 1985 (cont'd.)

472	04	495	88	518	11	541	11	564	81	587	49	610	00	706	18
473	04	496	88	519	11	542	11	565	81	588	00	611	00	707	18
474	02	497	88	520	04	543	11	566	81	589	30	612	00	708	18
475	02	498	88	521	43	544	11	567	81	590	30	613	00	709	18
476	02	499	88	522	43	545	83	568	81	591	30	614	00	710	18
477	02	500	88	523	43	546	83	569	81	592	30	615	00	711	18
478	06	501	04	524	43	547	83	570	81	593	30	616	00	712	18
479	06	502	02	525	53	548	83	571	81	594	28	617	00	713	18
480	06	503	04	526	99	549	83	572	81	595	28	618	00	714	18
481	06	504	04	527	99	550	81	573	81	596	00	619	00	715	18
482	06	505	13	528	49	551	81	574	76	597	00	620	00	716	18
483	06	506	13	529	49	552	81	575	27	598	00	621	00	717	18
484	04	507	11	530	08	553	81	576	27	599	00	622	00	718	18
485	04	508	11	531	96	554	81	577	11	600	16	623	00	719	18
486	90	509	88	532	96	555	81	578	08	601	14	624	00	720	18
487	90	510	88	533	96	556	81	579	08	602	00	625	00	721	18
488	90	511	88	534	96	557	81	580	19	603	00	626	00	722	18
489	88	512	88	535	94	558	81	581	99	604	00	700	18	723	18
490	88	513	88	536	94	559	81	582	99	605	00	701	18	724	28
491	88	514	86	537	94	560	81	583	99	606	00	702	18	725	18
492	88	515	86	538	94	561	81	584	97	607	00	703	18	726	18
493	88	516	04	539	94	562	81	585	51	608	00	704	18	727	10
494	88	517	04	540	11	563	81	586	78	609	00	705	18	728	14

*First three digits of the social security number are area numbers; second two digits are group numbers.

Group 00 is not a valid group -- it is for program purposes only.

ATTACHMENT D

Excerpt from
Validation and Screening Techniques for Social Security Numbers

VALIDATION OF SSN'S

Minimum information needed to validate an SSN is the person's name, sex, date of birth and the alleged SSN. Validation occurs only when the information on a current transaction exactly matches or can be reconciled with the information on the Alphident/Numident data bases or the microfilm subfiles of these systems. In certain circumstances, additional matching information is needed before validation can occur. If earnings are reported without an SSN or with an SSN or name that does not agree with these files and the correct SSN cannot be determined through internal screening operations, the employer or the worker is asked to furnish additional information to identify the record. The Internal Revenue Service (IRS) uses a similar system to validate SSN's of taxpayers.

MANUAL SCREENING OF DUPLICATE AND ORIGINAL SSN APPLICATIONS

The electronic screening operation to which every application is subjected is capable of processing roughly 85 percent of all applications input by field offices. Through a sophisticated series of screening grids, the computer makes a decision: is this applicant already represented in the Alphident data base? If the decision is yes, the previously assigned SSN is identified and a replacement card is prepared and mailed. If the decision is no, a number is assigned and a card is printed and mailed.

However, the decision-making capability of the system is deliberately limited because some applications have identifying information common to others or conditions exist which should receive a clerical review. These applications produce worksheets which are processed manually by OCRO.

Worksheets to be screened are checked against the Alphident Microfilm File and the Alphident Microfiche File, using the name and date of birth shown on the application. If an SSN is not located for the name and date of birth shown, another search is made using dates of birth somewhat different from the one given on the application. If an SSN is still not located, certain other variations are checked, including name at birth or on the signature line if different from the name in item 1; acceptable variations of common first names; dropping middle name shown; substituting different middle initials; substituting maiden surname for middle given name for married females; substituting initials only in place of complete given names; etc. Once a "possible" SSN is located, verification can be made immediately since full identifying information is available on the Alphident •files. See RM 00204.020 for procedures for handling "UTL" and "Investigate" items.

THE ALPHIDENT MICROFILM AND MICROFICHE FILES

The electronic Alphident file is updated daily. If an SSN holder loses the social security card within the first days after it was issued, the number can be located and verified electronically.

The Alphident Microfilm File is an alphabetical file based on the Russell Soundex coding system. It contains essentially the same information as the electronic file.

Because the Alphident Microfilm File is updated only every 3 months, each week an accretion file is prepared on microfiche. This file contains all SSN assignments and corrections to our records processed during the preceding 12 weeks. This file is referred to when there is reason to believe that there was a recent SSN action for an individual.

Each record entry on both the Alphident Microfilm and the Alphident Microfiche Files consists of the following:

DATA	POSITIONS
Blank	1
Soundex	2-5
Blank	6
Applicant's Surname	7-27
Applicant's Given Name	28-43
Applicant's Middle Name	44-45
Month of Birth	56-57
Blank	58
Day of Birth	59-60
Blank	61
Century of Birth	62
Year of Birth	63-64
Blanks	65-66
SSN	67-77
Blank	78
Mother's Surname	79-91
Mother's Given Name	92-102
Mother's Given Initial	103
Blank	104
Sex/Race	105-106
Blank	107
Father's Surname	108-120
Father's Given Name	121-131
Father's Middle Initial	132
Blank	133
City/County of Birth	134-140
State/Country of Birth	141-142
Blanks	143-144
Form/Entry	145-146
Blanks	147-148
Reference Number	149-159
Blank	160

COMMON NAMES IN THE ALPHIDENT FILE

There are over 360 million records in the Alphident File, representing over 277 million SSN's assigned. Many of the names in the file are the same or are very similar. This is why it is extremely important to get complete and accurate identifying information on original applications and on requests for duplicate SSN cards. It is equally important to obtain information that is consistent with that on the original application. Applicants who have lost their original cards should be questioned closely to find out if any of the information on the current application is now different from that which they showed on their original application.

The latest tabulation of common surnames in the SSN file was made in 1974. Some examples of the number of times a common name could appear in Alphident are given below.

NAME	NUMBER OF ITEMS IN ALPHIDENT
Smith	2,382,509
Johnso(n)	1,807,263
Willia(ms)(mson)	1,568,939
Brown	1,362,910
Jones	1,331,205
Miller	1,131,861
Davis	1,047,848
Martin(ez)(son)	1,046,297
Anders(on)	825,648
Wilson	787,825

THE RUSSELL SOUNDEX CODE

By using the Russell Soundex Code system, searching for possible SSN's on the Alphident film and fiche in OCRO is accomplished quickly.

Here are the basic rules for using the Soundex Code.

Use the first letter of the surname, then code the remaining letters as follows:

LETTERS	CODE SYMBOLS
BPFV	1
CGJKQSXZ	2
DT	3
L	4
MN	5
R	6

Vowels are not coded, nor are the letters W, H, and Y. Two successive letters with the same code numbers are coded only once.

Example:
"Mack" is coded M-200. The "a" is not coded since it is a vowel. "c" falls under code symbol 2. "k" also falls under code symbol 2, but is not used since two successive letters with the same code sumbol are coded only once. Since the complete Soundex Code must consist of the first letter of the name followed by three numbers, we add enough zeros to complete the 3-digit code.

Here are some other examples:

1. Snyder - S-536
2. Way - W-000
3. Bear - B-600
4. Brown - B-650

LIMITATIONS IN OCRO SCREENING FOR SSN's

When an applicant has indicated a previous SSN in item 10 of the SS-5 and the correct number cannot be found in the electronic or OCRO screening operations, the data are returned via form SSA-4310 to the district office. This is because studies show that many such applicants are mistaken in stating they previously applied for a number, and it is not worthwhile spending additional time on the case unless different information can be found. When the district office receives a form SSA-4310 from OCRO, it should recontact the applicant for any different information that may be useful in screening. See RM 00204.020 A.1. Take appropriate action, but do not return the SSA-4310 to OCRO.

Upon recontacting the applicant, the district office may discover that a married woman obtained her original SSN under a first husband's name, but is now applying for the duplicate in her second husband's name; that a man who calls himself "Winslow" obtained his number earlier in life as "Buddy;" or that Mr. Kline's record was set up originally under "Cline." There is also a possibility that the applicant may be able to locate the previously issued SSN on an old pay stub or by asking a present or a past employer. This new information may enable OCRO to locate the original SSN. If the applicant is unable to give any information different from what was previously given and is unable to locate the alleged number, the district office has no other choice but to request assignment of an original SSN. However, this should be done only as a last resort, particularly if the person has earnings under the original number which might not be credited when the SSN holder applies for benefits.

These facts point up the need for obtaining the most accurate information possible during the initial interview with the applicant, whether it be for an original or duplicate SSN card; otherwise, multiple numbers may result. Any reasonable assistance should be extended to the applicant to help find out definitely what the alleged prior SSN is. (See RM 00202.025 I.10.)

Source: "The Social Security Number," Program Operations Manual System, Part I, Chapter 00201.000, Section 00201.015, Social Security Administration.

EXACT MATCHING LISTS OF BUSINESSES: BLOCKING, SUBFIELD IDENTIFICATION, AND INFORMATION THEORY

William E. Winkler, Energy Information Administration

1. INTRODUCTION

The purpose of this paper is to present an evaluation of matching strategies for name and address files of businesses. In evaluating matching methods, we wish to minimize erroneous matches and nonmatches and the amount of manual review.

This work and previous work by various authors (Newcombe, Kennedy, Axford, and James, 1959; Newcombe and Kennedy, 1962; Newcombe, Smith, Howe, Mingay, Strugnell, and Abbatt, 1983; Coulter, 1977; Coulter and Mergerson, 1977; Rogot, Schwartz, O'Conor, and Olsen, 1983; Kelley, 1985) rely on matching strategies based on a theory of record linkage formalized by Fellegi and Sunter (1969) and first considered by Newcombe et al. (1959). The Fellegi-Sunter model provides an optimal means of obtaining weights associated with the quality of a match for pairs of records. Linked pairs (designated matches) and nonlinked pairs (designated nonmatches) receive high and low weights, respectively. Pairs designated for further manual followup receive weights between the sets of high and low weights.

Early work by Newcombe et al. (1959, 1962) showed the potential improvement (lower rates of erroneous matches and nonmatches and of manual followup) when weights were computed using surname and date of birth in comparison to when weights were computed using surname only. Coulter (1977) provided an example of the decrease in discriminating power as the probability of identifiers (such as surnames, first names, middle names, and place names) being misreported (transcribed inaccurately) and/or pairs of identifiers associated with individuals being different but accurately reported increases.

While the applied work referenced above involved files of individuals only, this paper provides an evaluation involving files of businesses. Matching using files of businesses is different from matching files of individuals because business files lack universally available and locatable identifiers such as surnames.

Matching consists of two stages. In the blocking stage, sort keys, such as SOUNDEX abbreviation of surname, are defined and used to create a subset of all pairs of records from files A and B that are to be merged. Records having the same sort key are in the same block and are considered during further review. Records outside blocks are designated as nonmatches. In the discrimination stage, surnames and other identifying characteristics are used in assigning a weight to each pair of records identified during the blocking stage.

With the exception of Newcombe et al. (1959, 1962), little work has been performed in evaluating how many erroneous nonmatches arise due to a given blocking strategy. The chief reason that little work has been performed is that identifying erroneous nonmatches due to blocking and accurately estimating error rates is difficult (Fellegi and Sunter, 1969; Winkler, 1984a,b).

The key to identifying difficulties in blocking files of businesses is having a data base in which all matches are identified and which is representative of problems in many business files. In section 2, the construction of such a data base from 11 Energy Information Administration (EIA) and 47 State and industry files is described. Section 2 also contains a summary of the Fellegi-Sunter model and the criteria used in evaluating competing matching strategies.

Section 3 is divided into two parts. The first part contains results obtained by multiple blocking strategies using a procedure in which the numbers of erroneous nonmatches and matches are minimized under a predetermined bound on the number of pairs to be passed on to the discrimination stage (for related work see Kelley, 1985). The results are related to results obtained during the discrimination stage and build on earlier work of Winkler (1984a, 1984b).

In the second part, the main results of the discrimination stage are presented. The effects of improved spelling standardization procedures and identification of additional comparative subfields are highlighted. Although the deleterious effect of poor spelling standardization is covered by the Fellegi-Sunter theory and presented in the simulation results of Coulter (1977), no concrete examples have previously been presented.

The second part also contains results on the variation of cutoff weights and misclassification and nonclassification rates during the discrimination stage. The results are based on small samples used for calibration and obtained using multiple imputation (Rubin, 1978; Herzog and Rubin, 1983) and bootstrap imputation (Efron, 1979; Efron and Gong, 1983). Fellegi and Sunter (1969, p. 1191) indicate that results based on samples are unreliable.

Finally, the second part presents results addressing the strong independence assumptions necessary under the Fellegi-Sunter model and conditioning techniques that can be used in improving matching performance in some situations when direct application of the Fellegi-Sunter model yields high misclassification and/or nonclassification rates. The investigation of independence uses the hierarchical approach of contingency table analysis (Bishop, Fienberg, and Holland, 1975). The conditioning argument uses a steepest ascent approach (Cochran and Cox, 1957).

Section 4 contains a summary and further discussion of the results and problems for future research.

2. EMPIRICAL DATA BASE, METHODS, AND EVALUATION CRITERIA

This paper´s approach to developing more effective matching strategies involves:

1. constructing an empirical data base for testing procedures;
2. employing the Fellegi-Sunter model of record linkage;
3. defining evaluation criteria; and
4. refining procedures in response to empirical results.

A suitable data base should have all duplicates identified and connected to their respective parents (records used for mailing purposes) and present problems that are representative of similar data files (in this case, files of businesses). The identification of all duplicates allows determination of erroneous nonmatches during the blocking stage. Evaluation criteria should be such that they are suitable for adoption by others performing research in matching methodologies.

2.1. Creation of a Suitable Empirical Data Base

The empirical data base consists of 66,000 records of sellers of petroleum products. It was constructed from 11 EIA lists and 47 State and industry lists containing 176,000 records. Easily identified duplicates having essentially similar NAME and ADDRESS fields were deleted when the melded file was reduced from 176,000 to 66,000 records.

The data base contains 54,850 records identified as headquarters or parents (records used for mailing purposes); 3,050 records identified as duplicates (records having names and addresses similar to their parents´); and 8,511 records identified as associates (records such as subsidiaries and branches that have names and/or addresses different from their parents´).

Duplicates were identified primarily through elementary computer-assisted techniques (see Winkler, 1984a); associates were identified through surveying and call-backs. Our evaluation will only consider how well various strategies perform in matching duplicates with headquarters. The presence of unidentified associates, however, can cause falsely higher error rates (see section 2.3.1).

2.1.1. General Applicability of Results

Procedures developed for dealing with problems in the main empirical data base would be generally applicable to most EIA systems because the data base:

1. is larger than any other master frame file in EIA;
2. is involved with retail sales-- such frames are often more difficult to work with than files of individuals or files of headquarter addresses of large corporations; and
3. had greater formatting and spelling standardization difficulties-- it was constructed from many more sources than any other EIA frame.

Because the main empirical date base is constructed from many different lists and contains many records associated with retailers, results should be representative of the difficulties encountered with similarly constructed, non-energy files of businesses.

2.1.2. Improved Spelling Standardization

The original spelling standardization software contained two basic loops. The first replaced most punctuation with blanks and deleted multiple blanks within a field. The second used lookup tables to replace a given spelling of a word with a standardized spelling or abbreviation. Blanks were generally used to delimit words within fields.

Spelling standarization software was updated in two ways. First, the logic of the processing was enhanced to cause changes in character strings that are not easily updated because they contain embedded punctuation or blanks. For instance, "´S" is replaced by "S" and "MC NEELY" by "MCNEELY."

Second, standardization tables were updated with a very large number of spelling variations of words such as ´COMPANY,´ ´DISTRIBUTOR,´ ´SERVICE,´ and ´CORPORATION.´ The key to systematically identifying such spelling variations was a program that created an alphabetic listing and frequency count of every word in a prespecified field such as NAME or STREET ADDRESS. As more than 90 percent of keypunch errors occur after the first character (see e.g., Pollock and Zamora, 1984), most spelling variations of commonly occurring words in the empirical data base have probably been identified.

2.1.3. Identification of Subfields

The identification of subfields was done in two stages. In the first, ZIPSTAN software (U.S. Dept. of Commerce, 1978b) was used to process the STREET ADDRESS field. Although the Census Bureau uses a UNIVAC computer system, we were able to obtain an unsupported version of ZIPSTAN that had been created for use on IBM systems.

The basic idea of ZIPSTAN was to identify key subfields of the STREET ADDRESS field for files of individuals. Although ZIPSTAN assumes that the street address begins with a numeric word, which is the usual situation in the files of individuals for which ZIPSTAN was designed, it is able to process other types of street address subfields that typically occur in files of establishments or businesses.

Although ZIPSTAN provided warning messages for 18 percent of the 66,410 records in the empirical data base, it was still helpful for most cases. Warning messages consisted of ´MISSING STATE NAMES´ (records associated with non-US postal addresses), ´PLACE NAMES CONVERTED´ (minor conversion of the city field), ´STREET NAMES CONVERTED´ (minor conversion of the street name), ´SYNTAX CONVERSION´ (conversion of unacceptable patterns of word characteristics), and ´POST OFFICE BOXES´ (containing PO BOX).

The following examples show some representative EIA records before and after ZIPSTAN processing.

Before ZIPSTAN

1. EXCH ST
2. HWY 17 S
3. 1435 BANK OF THE
4. 2837 ROE BLVD
5. MAIN & ELM STS
6. CORNER OF MAIN & ELM
7. 100 N COURT SQ
8. 100 COURT SQ SUITE 167
9. 2589 WILLIAMS DR APT 6
10. 15 RAILROAD AVE
11. 2ND AVE HWY 10 W
12. MAIN ST
13. 184 N DU PONT PKWY
14. 1230 16TH ST
15. BOX 480

After ZIPSTAN

No.	House No.	Prefixes 1	Prefixes 2	Street Name	Suffixes 1	Suffixes 2	Unit
1.				EXCH	ST		
2.		HW		17TH	S		
3.	1435			BANK OF THE			
4.	2837			ROE	BL		
5.				MAIN ELM STS			
6.				CORNER OF MAIN ELM			
7.	100	N		COURT	SQ		
8.	100	CT	SQ	*** NO NAME ***			RM 167
9.	2589			WILLIAMS	DR		AP 6
10.	15			RAILROAD	AV		
11.				2ND	AV	HW	10
12.				MAIN	ST		
13.	184	N		DU PONT	PW		
14.	1230			16TH	ST		
15.	480			*PO BOX*			

ZIPSTAN is able to identify accurately subfields in 13 of 15 cases. The two exceptions are cases 2 and 8. In case 2, ´HWY´ is moved to a prefix position and ´17´ is placed in the STREET NAME position. In case 8, ´COURT,´ the street name, is placed in a prefix location.

Although ZIPSTAN accurately identifies the subfields associated with intersections (cases 5, 6, and 11), such identification may not allow accurate delineation of duplicates in comparisons of various lists. Some lists may contain STREET ADDRESSes in the following forms, none of which can be readily comparable with the forms in examples 5, 6, and 11.

5. 34 Main St
5. Elm and Main Streets
11. Hwy 10 W
11. 7456 Richmond Hwy

In the second stage of subfield identification, the following words in the NAME field were identified:

KEYWORD1	Largest word in NAME field
KEYWORD2	2nd largest word in NAME field (ties broken by alpha sort)
CON	Concatenation of initials

The above three subfields were used for comparison purposes because the NAME field in lists of businesses generally does not contain words such as SURNAME and FIRST NAME that are present in files of individuals. Based on a sample of 1000 records, an upper bound of 27 percent at the 95 percent confidence level is placed on the number of records containing a word that could be identified as SURNAME.

The identification of SURNAMEs was not performed for three reasons: (1) it is difficult to develop software that accurately identifies records that contain SURNAME (see U.S. Dept. of Agriculture, 1979); (2) it is difficult develop software to identify SURNAMES within the NAME field (e.g., PAUL ROBERT or ROBERT PAUL- which is the SURNAME?); and (3) the small number of records to be compared and containing surnames was not sufficient to justify such a development effort.

The following provides examples of legitimate variations associated with NAME field of one company:

J K Smith Co
Smith Jonathon K
Smith Fuel Service Co
J K Smith Exxon Fuel Service
J K S Fuel

Fellegi and Sunter (1969, pp. 1193-1194) provide an explicit theoretical model for how much such legitimate spelling variations decrease the accuracy with which matches and nonmatches are delineated. Coulter (1977) provides an empirical example of the decrease based on a simulation.

Identifying and comparing the largest words in the NAME field are only performed after spelling standardization and/or abbreviation so´that the chance of designating large words with little distinguishing power is minimized.

For instance, if a character string such as ´DISTRIBUTOR´ appeared in the name field, it would likely be the longest word. Replacing the various spellings of ´DISTRIBUTOR´ with an abbreviation such as ´DSTR´ either allows it to be deleted so that it is not considered by the keyword-identification program or allows longer words with possibly more distinguishing power to be identified.

Although methods of identifying subfields might be considered results, we are primarily concerned with how their identification affects the efficacy of various matching procedures. Consequently, the identification can be considered a preprocessing step (see e.g., Winkler, 1985) that is used in creating the data base used in evaluations.

2.1.4. Completeness of Identification of Duplicates

It is likely that few, if any, additional erroneous nonmatches of duplicates are present in the empirical data base for three reasons. First, no additional duplicates were identified in the set of headquarters records during a manual review of all 1,500 records in a random sample of 3-digit ZIP codes. Second, no additional duplicates were identified during a review of a sample of 20 pages (each containing 60 records) in a listing that was ordered alphabetically using the NAME field. Third, no additional duplicates were identified during the discrimination stage (section 3.2).

Without further manual followup, it is impossible to determine how many unidentified associate records are in the set of headquarters records. It is unlikely that surveying and callbacks--because they were first-time efforts--would have been able to identify them all.

Even if more associates are identified, the results of matching duplicates against headquarters will not be seriously affected. The main effect of identifying more associates will be to lower the estimated rates of erroneous matches. Some duplicates are now matched to headquarters that are not identified as their parent and that are actually associates of the duplicates' parents. Each such match is presently counted as an erroneous match.

2.2. Methods

2.2.1. The Formal Probabilistic Model

The Fellegi-Sunter model (1969) uses an information-theoretic approach embodying principles first used in practice by Newcombe (Newcombe et al., 1959). For a review of existing techniques and their relationship to classical information theory see Kirkendall (1985).

In the Fellegi-Sunter model, agreements on characteristics such as SURNAME or ZIP code are assumed to be more common among truly matched pairs than among erroneously matched or unblocked pairs. In practice, specific binit weights of agreement (or disagreement) are computed by,

$$W = \log_2 A/B$$

where

A= the proportion of a particular agreement (or disagreement) defined as specifically as one wishes among matched pairs, and

B= the corresponding proportion of the same agreement (or disagreement) among pairs that are rejected as matches.

The following table will help us to understand more specifically the computation of weights.

Table 1: Counts of True State of Affairs

Specified Characteristic	Match	Nonmatch
Agree	a	b
Disagree	c	d

If we wish to compute the weight associated with agreement on a specified characteristic, then we take A=a/(a+c) and B=b/(b+d); for disagreement, we take A=c/(a+c) and B=d/(b+d).

For each detailed comparison of a pair of records, the weights for appropriate agreements and disagreements are added together, and the total weight, TWT, is used to indicate the degree of assurance that the pair relates to the same entity. The procedure assumes that weights associated with individual agreements or disagreements are uncorrelated with each other (at least conditionally, see e.g., Fellegi and Sunter, 1969, p. 1190).

Cutoffs UPPER and LOWER are chosen (using empirical knowledge or educated guesses) and the following decision rule is used:

If TWT > UPPER, then designate pair as a match.

If LOWER <= TWT <= UPPER, then hold for manual review.

If TWT < LOWER, then designate pair as a nonmatch.

Given fixed upper bounds on the percentages of erroneous nonmatches having TWT < LOWER and of erroneous matches having TWT > UPPER, Fellegi and Sunter (1969, p. 1187) show that their procedure is optimal in the sense that it minimizes the size of the manual review region.

In some cases, either looking at disjoint subsets of the set of blocked pairs and/or increasing or decreasing individual weights used in computing the total weight, TWT, can improve the efficacy of the above decision rule. For instance, among a set of records that are blocked into pairs using the first six characters of the STREET field, individual weights associated with agreements and disagreements on characteristics of the NAME field might be increased and decreased, respectively.

A procedure that uses individual weights, that have been varied in order to achieve greater accuracy in the set of pairs designated as matches and nonmatches and/or a reduction in the set of records held for manual review, will be referred to as a modified information-theoretic procedure. An unmodified procedure will be referred to as the basic information-theoretic procedure.

2.2.2. Specific Weight Computation

In addition to individual weights computed using the subfields HOUSE NUMBER, PREFIX, STREET NAME, SUFFIX, UNIT DESIGNATOR, KEYWORD1, KEYWORD2, and CO given in section 2.1.3, the following subfields were used in computing individual weights:

Field	Subfield Columns	Designated as
NAME	1-4,5-10,11-20,21-30	N1,N2,N3,N4
STREET	1-6,7-15,16-30	S1,S2,S3
ZIP	1-3,4-5	Z1,Z2
CITY	1-5,6-10,11-15	C1,C2,C3
STATE	1-2	
TELEPHONE	1-3,4-6,7-10	T1,T2,T3
WL-NAME 1/	1-4,5-10,11-20,21-30	W1,W2,W3,W4

1/ Sort words in NAME field by decreasing order of wordlength. Break ties with alpha sort.

Generally, corresponding subfields were used in computing individual weights. The exceptions were comparisons of the first and second keywords (section 2.1.3) in the NAME field.

It is important to note that if any weight associated with a given SORT KEY, say TELEPHONE, used in blocking is computed only for records within the subset of pairs having the SORT KEY agreeing, then the comparison has no discriminating power and the resulting weight is zero. If, however, a weight is computed for a comparison of a SORT KEY within a subset of pairs which do not all agree on the SORT KEY, then the weight could be nonzero. Also, it is intuitive that some of the comparisons, say of the above defined subfields of the NAME and KEYWORDs (section 2.1.3) may not be independent.

2.2.3. Variances

As the truth and falsehood of matches in the set of blocked pairs were known for the evaluation files, estimated error rates and their variances were obtained using multiple samples.

The basic procedure was to draw samples of equal size, compute cutoff weights using each sample (based on at most 2 percent of nonmatches being classified as matches and at most 3 percent of matches being classified as nonmatches), use each pair of cutoff weights on the entire data base to determine overall error rates, and compute the variances of the cutoff weights and the overall error rates over the set of samples.

The multiple imputation procedure of Rubin (1978) has been used for evaluating the effects of different methods of imputing for missing data but is applicable in our situation. Multiple imputation entails obtaining several estimates using different samples and then computing the mean and variance over samples. In using Rubin´s procedure, we sample without replacement.

The key difference from Efron´s bootstrap is that sampling is performed with replacement. Our application corresponds almost exactly to the first example in the paper of Efron and Gong (1983).

2.2.4. The Independence Assumption

Fellegi and Sunter (1969, pp. 1189-90) state that the independence assumption for the comparisons of information contained in different subfields is crucial to their theory but that the independence assumption may not be crucial in practice. They note that obtaining total weights having a probabilistic interpretation only necessitates that comparisons be conditionally independent. The conditioning must be consistent with the way total weights are computed.

There are several practical difficulties with testing their independence assumption. First, it must be tested separately for matches and nonmatches. Newcombe and Kennedy (1962) provide a method of approximating the weights for nonmatches and show that accurately approximating the weights for matches is difficult. The chief reason is that the number of nonmatches is close to the number of pairs in the cross product of two files A and B while matches represent a relatively small subset (of all pairs) having specific characteristics.

Second, the weights of nonmatches and matches may vary substantially depending on what blocking criteria are used. If, say, four independent criteria are used, then it might be necessary to examine as many as 15 (2**4-1) mutually exclusive subsets of the set of blocked pairs (see sections 3.1 and 3.2).

Third, the collection of the information necessary for contingency table analyses is difficult because we have no strong control over sampling design (Bishop, Fienberg, and Holland, 1975, pp. 36-39). Even with moderately large samples, some of the subsets determined by blocking criteria may be too small for adequate analysis of the conditional independence of two variables given two or more variables because of the number of marginal constraints that are zero (see section 3.2.8).

Fourth, if many different subfields and/or different means of comparing them are considered (we will consider 30; Newcombe and Kennedy, (1962, p. 566), considered 200), then modelling the conditional relationships using contingency table techniques (Bishop, Fienberg, and Holland, 1975) can be cumbersome.

Even if dependencies occur, it may be possible to vary weights associated with individual comparisons (i.e., steepest ascent, see e.g., Cochran and Cox, 1957, pp. 357-369) to determine whether the efficacy of the overall weighting procedures can be improved. Our specific steepest ascent method generally involved choosing a few individual weights in disjoint subsets determined by blocking criteria (sections 3.1 and 3.2) and varying them by +/- 0.5.

It is important to note that modifications to individual weights may be heavily dependent on the subsets determined by the blocking criteria.

2.3. Criteria for Evaluation

2.3.1. Type I and II Errors

A Type I error is an erroneous nonmatch and a Type II error is an erroneous match. The Type I error rate is U/D*100 where U is the number of erroneous nonmatches and D is the number of matches. The Type II error rate is F/M*100 where M is the number of pairs designated as matches and F is the number of erroneous matches.

As duplicates unmatched during the blocking stage are considerably more difficult to identify than false matches during the discrimination stage, the primary emphasis in developing a new strategy was minimizing Type I errors during the blocking stage before minimizing Type II and Type I errors during the discrimination stage.

It is important to note that if a pair of files has no erroneous nonmatches, then any matching strategy applied will yield either no pairs during the blocking stage or a Type I error rate of 0 percent and a Type II error rate of 100 percent. Because the empirical data base is relatively free of duplicates (as a result of reducing the empirical database from 176,000 to 66,000 records), application of any matching strategy will produce relatively high Type I error rates during the blocking stage.

As we are primarily concerned with evaluating methodologies for accurately matching pairs that are not readily matched using elementary comparisons (e.g., having major portions of key fields agreeing exactly), the data base of 66,000 records is more suitable for use than the original set of 176,000 records.

2.3.2. Overall Rate of Duplication

The number of erroneous nonmatches as a percentage of the total number of records in a file is also an important evaluation criteria. We define the overall rate of duplication as Q/(X+Q)*100 where Q is the number of erroneous nonmatches and X is the number of parent records.

This additional evaluation criteria is important because the Type II error rate criteria will not provide a measure of how free of duplicates a file is. The Type II error rate does not work well because, as the number of matches, D, in a file decreases, the Type I error rate (U/D*100, where U is the number of erroneous nonmatches) will necessarily increase.

In the analysis of the empirical data base, D is held constant so that the comparative advantages of various strategies can be assessed using Type I error rates. The overall rate of duplication will not work well for these comparative evaluations because it is too dependent on the number of parent records, X, which does not change. That is, if U1 and U2 are the numbers of erroneous nonmatches under two matching strategies and U1<U2<<X, then U1/(U1+X) and U2/(U2+X) are approximately equal.

2.3.3. Amount of Manual Review

The amount of manual review is a critical feature in any matching procedure because manual review is both time-consuming and expensive. If one procedure requires one half as much manual review as another, yields Type I error rates that are only somewhat higher than the other, and yields similar rates of erroneous nonmatches (section 2.3.2), then there is strong justification for adopting the procedure requiring less manual review.

3. RESULTS USING THE EMPIRICAL DATA BASE

Results of the empirical analyses for the blocking stage and the discrimination stage are presented in sections 3.1 and 3.2 respectively.

3.1. Comparison of Sets of Blocking Strategies

The following five criteria were used for blocking files into sets of linked pairs used in the discrimination stage. The set of five criteria were developed by comparing a large number of criteria. If the upper bound on the overall rate of erroneous matches during the blocking stage is set at 65 percent, then this set of five gave the largest overall reduction in erroneous nonmatches (see Winkler, 1984a).

BLOCKING CRITERIA

1. 3 digits ZIP, 4 characters NAME
2. 5 digits ZIP, 6 characters STREET
3. 10 digits TELEPHONE
4. Word length sort NAME field, then use 1. *
5. 10 characters NAME

* This criterion also has a deletion stage which prevents matching on commonly occurring words such as 'OIL,' 'FUEL,' 'CORP,' and 'DISTRIBUTOR.'

3.1.1. Type I and II Error Rates by Individual Blocking Criteria

Table 2 presents counts and rates of matches, erroneous matches, and erroneous nonmatches for each of the five matching criteria given above.

As we can see, no single criterion provides a significant reduction in the rate of erroneous nonmatches. The best is criterion 4 (wordlength sort) which leaves 702 (23 percent) duplicates unlinked. The reason criteron 4 works best is that the NAME field does not have subfields (generally words) that are in fixed order or in fixed locations. Consequently, criterion 4 links NAME fields from headquarters and duplicates having the following form:

John K Smith
Smith J K Co

Criterion 3 (TELEPHONE) provides the lowest rate 8.7 percent (186/(186+1952)) of erroneous matches and the second best rate 34.7 percent (1057/3050) of erroneous nonmatches. Criterion 5 (10 characters of the NAME) provides both the worst rate of erroneous matches, 58.6 percent (1259/1259+889)), and the worst rate of erroneous nonmatches, 63.3 percent (1932/3050).

Table 2: Rates of Matches, Erroneous Matches, and Erroneous Nonmatches by Blocking Criteria

Criterion	Link with Correct Parent 1/	Link with Wrong Parent	Not Linked 2/	Actual Number of Matches
1	1460 (66.8)	727	1387 (45.5)	3050
2	1894 (82.5)	401	1073 (35.2)	3050
3	1952 (91.3)	186	1057 (34.7)	3050
4	2261 (80.3)	555	702 (23.0)	3050
5	763 (14.4)	4534	1902 (62.4)	3050

1/ Type II error rates are in parentheses.
2/ Type I error rates are in parentheses.

3.1.2. Comparison of Sets of Criteria

In comparing subsets of the five blocking criteria, the primary concern is in reducing the number of erroneous nonmatches. The number of matches and erroneous matches in the set of pairs created in the blocking stage is dealt with primarily during the discrimination stage.

The comparison takes the form of considering the incremental reduction in the number of erroneous nonmatches as each individual criteria is added. Although criteria 3 and 4 perform best on the empirical data base, they are considered later than criteria 1 and 2.

Criteria 1 and 2 are applicable to all EIA files because all of them have identified NAME and ADDRESS fields. As many non-EIA source lists used in updating do not contain telephone numbers, criterion 3 is not applicable to them. As a number of EIA lists have consistently formatted NAME fields, criterion 4 will yield little, if any, incremental reductions in the number of erroneous matches during the blocking stage.

Table 3: Incremental Decrease in Erroneous Nonmatches and Incremental Increase in Matches and Erroneous Matches by Sets of Blocking Criteria

Set of Criteria Used	Rate of Erroneous Nonmatches	Erroneous Nonmatches/ Incremental Decrease	Matches/ Incremental Increase	Erroneous Matches/ Incremental Increase
1	45.5	1387/ NA	1460/ NA	727/ NA
1,2	15.1	460/927	2495/1035	1109/ 289
1,2,3	3.7	112/348	2908/ 413	1233/ 124
1,2,3,4	1.3	39/ 73	2991/ 83	1494/ 261
1,2,3,4,5	0.7	22/ 17	3007/ 16	5857/4363

NA- not applicable.

3.1.3. The Preferred Set of Blocking Criteria

The preferred set of blocking criteria are criteria 1, 2, 3, and 4. Criterion 5 (10 characters of the NAME) was considered because it yielded the greatest reduction in erroneous nonmatches of any fifth blocking criteria while keeping the overall percentage of erroneous matches below 65 percent.

Criterion 5, however, is not suitable for inclusion because it incrementally adds 16 matches and 4363 erroneous matches while reducing the number of erroneous nonmatches from 39 to 22. As the discrimination stage (section 3.2) delineates matches and nonmatches with an error rate of 3 percent and 99.6 (4363/4379) of the incrementally-added pairs are false, inclusion of criterion 5 would yield an overall increase in the number of erroneous nonmatches.

Blocking 3050 duplicates with 54,850 parents using the preferred set of blocking criteria yielded 4485 pairs (2991 matches and 1494 nonmatches) for consideration during the discrimination stage.

It is important to note that the 39 matches not identified during the blocking stage are never again considered. Erroneous matches created during the blocking stage are considered during the discrimination stage and still can be correctly designated. These reasons led to our emphasis on minimization of Type I errors during the blocking stage prior to minimization of Type I and II errors during the blocking stage.

3.2. Discrimination

The discrimination stage was divided into two parts: (1) a part in which 2240 pairs were designated as matches using an ad hoc decision rule and (2) a discrimination stage in which the remaining 2245 pairs were designated as either matches, erroneous matches, or candidates for manual review.

The ad hoc decision rule generally consisted of designating those pairs as matches that had been connected by two or more blocking criteria. The exceptions were records connected by 1 and 4, only (NAME and WL-NAME), and 2 and 3, only (STREET and TELEPHONE). Slightly more than 98 percent of the 2240 records designated as matches were actually matches.

Prior to use in the information-theoretic discrimination procedure, the 2245 remaining pairs were further divided into four mutually exclusive classes using the preferred blocking criteria (section 3.1.3):

Class 1 (1021 records): Linked by 1, only, and by 1 and 4, only.
Class 2 (624 records): Linked by 2, only, and by 2 and 3, only.
Class 3 (256 records): Linked by 3, only.
Class 4 (344 records): Linked by 4, only.

3.2.1. Overall Results

Table 4 presents a summary of results obtained during the discrimination stage. It shows that 2148 (96 percent) of 2245 records are classified as matches or nonmatches and that only 3 percent (68/2148) of the classified records are misclassified. Results are based on using the entire data set for calibration (i.e., obtaining cutoff weights) and evaluation. Variance results (section 3.2.6) based on 25 different samples used for calibration yield cutoff weights and error rates that are consistent with results in Table 4.

Two observations are that the cutoff weights vary substantially across classes and that 100 percent of the records in classes 2 and 4 can be classified. The varying cutoff weights indicate that cutoff weights may vary with different types of address lists. Thus, new calibration information may be needed for each new file encounted. Calibration information is based on knowing the actual truth and falsehood of matches within a representative set of blocked pairs.

Table 4: Results from Using a Modified Information-Theoretic Model for Delineating Matches and Erroneous Matches (3 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassed as		Total Classed as		Total Classed	Total Records
	LOWER	UPPER	Non-Match	Match	Non-Match	Match		
1	4.5	7.5	28	8	692	274	966	1021
2	2.5	2.5	5	3	379	245	624	624
3	-0.5	4.5	5	6	104	110	214	256
4	8.5	8.5	9	4	266	78	344	344
Totals			47	21	1441	707	2148	2245

The largest group of misclassified records are those erroneous matches that have the same address and phone number as the headquarters' records. For example:

(a) Apex Oil 222 Columbia St NE Salem
OR 97303 503/588-0455
Jones Co 222 Columbia St N E Salem
OR 97303 503/588-0455
(b) A A Oil Main St Smallsville TX
77103 713/643-2121
Smith J K Co Main St Smallsville TX
77103 713/643-2121

Example (a) represents two different companies located in the same office building. Example (b) represents two different fuel oil dealers, one of which has gone out-of-business.

Misclassified matches (erroneous nonmatches) generally had typographical differences or missing data in a number of subfields, as in the examples below:

(c)	Smith Oil Hardsburg	W 31st St N Church St PA 18207	 713/643-2121
	Smith J K Hardsburg	N Church St PA 18207	 missing
(d)	Mcneely R MPLS	3312-14 Harris Ave MN 55246	 612/929-6677
	R Mcden Neely St Louis Par	3312 Harris Ave MN 55246	 612/929-6677

Example (c) has a minor variation in the NAME field, a major variation in the STREET field, and a missing TELEPHONE field. Example (d) has major variations in the NAME field and CITY fields and a minor variation in the STREET field.

3.2.2. Improvement Due to New Spelling Standardization

The improvement due to the new spelling standardization was quite minor as the results in Figures 1 and 2 show. Figures 1 and 2 represent plots of the numbers of matches and nonmatches against total weight using the early and new spelling standardizations, respectively.

The results are only shown for Class 2 (section 3.2 and section 3.1.3) because records blocked using STREET ADDRESS only or STREET ADDRESS and TELEPHONE only are intuitively among the most difficult to work with (see examples in section 3.2.1). Both figures will be compared with other figures corresponding to Class 2 that appear in sections 3.2.2, 3.2.3, and 3.2.4. Although characteristic results for other classes will be mentioned, no graphs will be presented for them.

Figures 1 and 2 show the classic patterns in matches and nonmatches (Newcombe et al., 1959; Newcombe et al., 1983; Rogot et al., 1983). In both figures, the curves of matches almost entirely overlap with the curves of nonmatches. As the distinguishing power of the weighting scheme improves, the curves move apart.

FIGURE 1: Total Weight Versus Counts of Matches and Nonmatches Prior to New Spelling Standardization Prior to Identification of Subfields

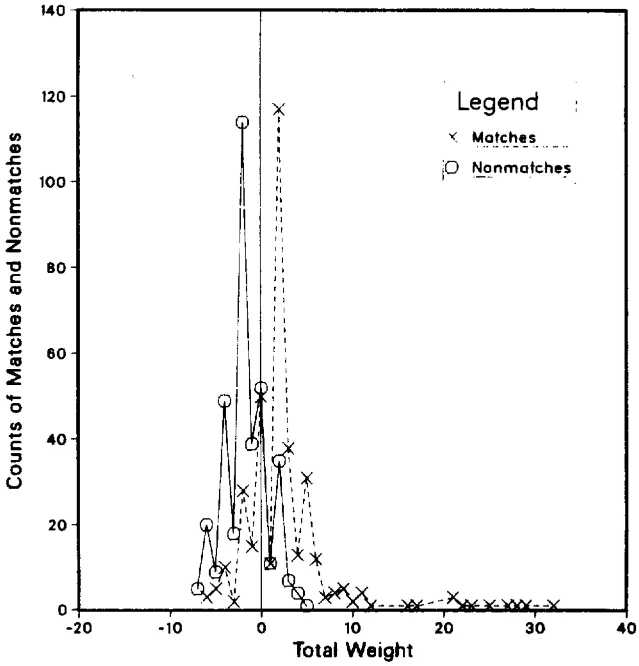
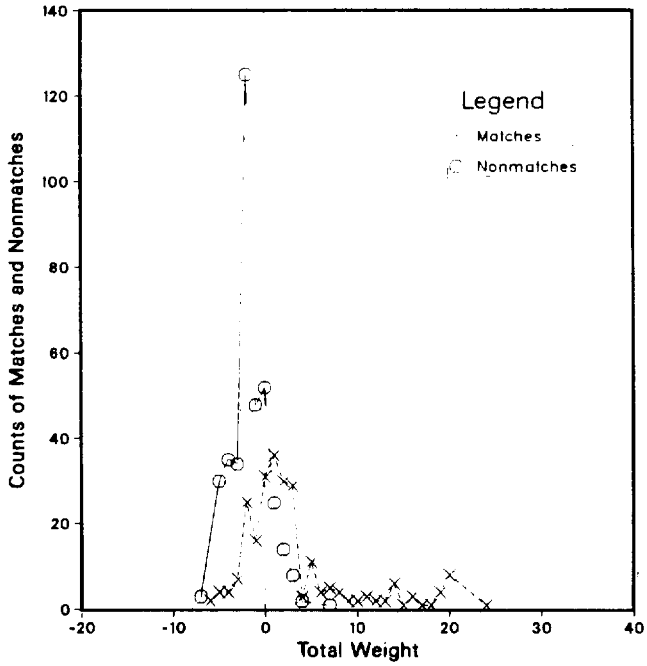


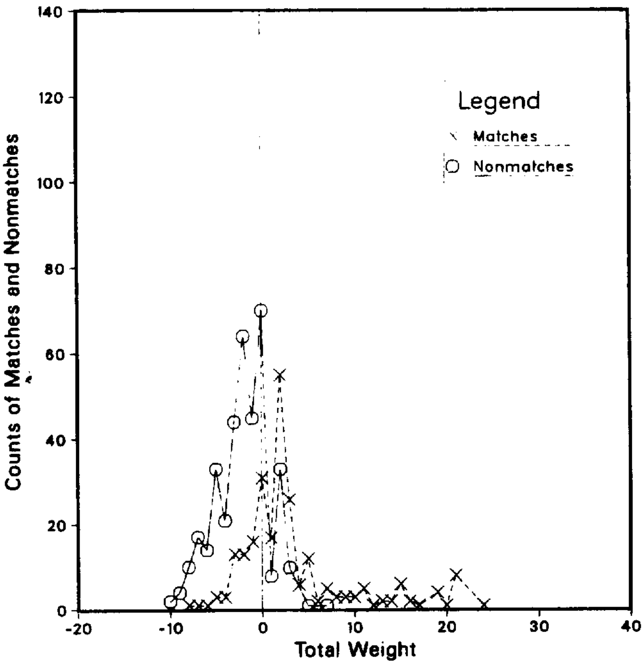
FIGURE 2: Total Weight Versus Counts of Matches and Nonmatches After New Spelling Standardization Prior to Identification of Subfields



3.2.3. Improvement Due to Address Subfield Identification

Figure 3 is a plot of the numbers of matches

FIGURE 3: Total Weight Versus Counts of Matches and Nonmatches After New Spelling Standardization Address Subfield Identification



and nonmatches against total weight when the new spelling standardization and address subfield identification (section 2.1.3) is used. Comparison with Figure 2 shows that the subfield identification yields a moderate improvement (i.e., the curves of matches and nonmatches overlap less.)

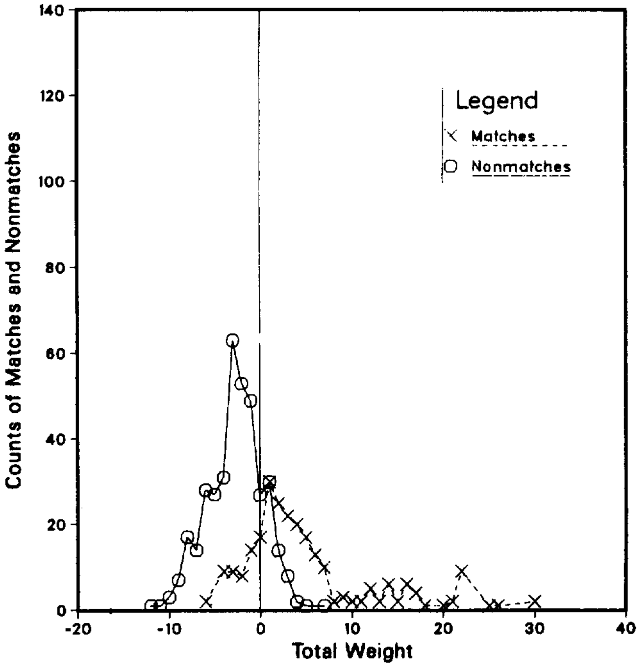
Although not shown in this paper, examination of similar sets of plots for other classes, particularly those blocked using the NAME field, show less improvement when additional weights obtained using the ADDRESS subfields are used.

3.2.4. Improvement Due to Name Subfield Identification

Figure 4 is a plot of the numbers of matches and nonmatches against total weight when the new spelling standardization and name and address subfield identification are used (see section 2.1.3 for a list of the subfields). Comparison with Figure 3 shows that the NAME subfield identification yields little, if any, improvement.

Although not shown in this paper, examination of similar sets of plots for other classes, particularly those blocked using the NAME field, show greater improvement when additional weights obtained using the NAME subfields are used.

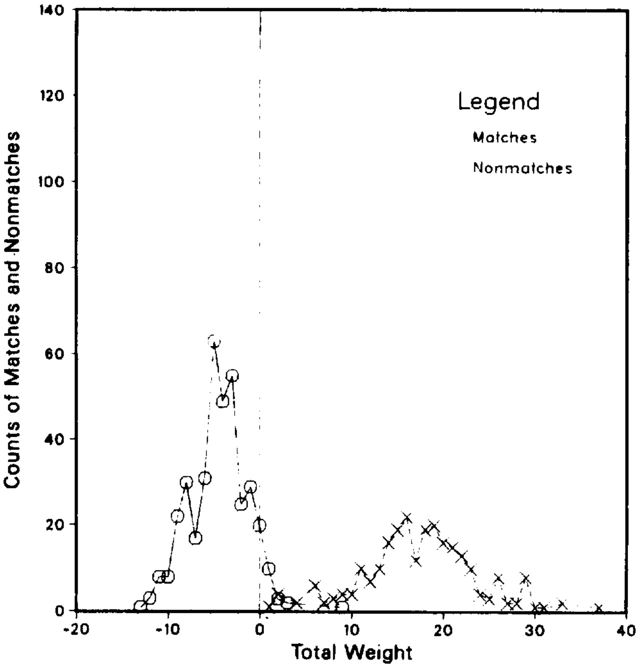
FIGURE 4: Total Weight Versus Counts of Matches and Nonmatches After New Spelling Standardization Name and Address Subfield Identification



3.2.5. Improvement Due to Conditioning

Figure 5 is a plot of the numbers of matches and nonmatches against total weight when a special conditioning (see section 2.2 and section 3.2.8) procedure in addition to the new spelling standardization and name and address subfield identification is used. Comparison with Figure 4 shows that the conditioning yields a substantial improvement in Class 2. Other classes (not shown) show slight improvements.

FIGURE 5: Total Weight Versus Counts of Matches and Nonmatches Name and Address Subfield Identification Conditioning



Comparison of Figure 5 with Figures 1 or 2 show the significant improvements obtained using the modified information-theoretic model that includes all enhancements.

Table 5 shows the results from using the basic information-theoretic model that are comparable to the results in Table 4. The only difference is that a modified information-theoretic procedure is used in obtaining Table 4 results. Overall comparison shows that the modified information-theoretic procedure performs better than the basic information-theoretic procedure.

Specifically, comparison of the two tables shows that the total number of records classified rises from 1526 (out of 2245) to 2148 while the overall misclassification rate falls from 5 percent to 3 percent.

Comparison of Tables 4 and 5 also shows that the main difference in the modified and basic procedures is that the modified procedure allows classification of all 624 records in class 2 while the basic procedure allows classification of only 215.

Table 5: Results from Using an Information-Theoretic Model for Delineating Matches and Erroneous Matches (5 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassed as		Total Classed as		Total Classed	Total Records
	LOWER	UPPER	Non-Match	Match	Non-Match	Match		
1	0.5	6.5	39	14	674	264	938	1021
2	-4.5	3.5	2	4	100	115	215	624
3	-4.5	6.5	2	1	55	42	97	256
4	2.5	11.5	11	2	254	46	300	344
Totals			54	21	1055	471	1526	2245

3.2.6. Variances

Tables 6, 7, and 8 present estimates and their coefficients of variation obtained using 25 calibration samples and Rubin's multiple imputation technique. For each calibration sample, the sample sizes in Classes 1, 2, 3, and 4 were 240, 200, 120, and 160, respectively. Cutoff weights and misclassification rates were obtained for each sample. Estimates are the average cutoff weights and average misclassification rates over 25 replications (samples). Variances of the estimates are over 25 replications.

Overall, the results indicate that the estimated cutoff weights and misclassification rates vary significantly from calibration sample to calibration sample. The variances are functions of both the sample sizes on each replication and the number of replications. When the number of replications was held at 25 and the sample sizes decreased to 120, 100, 80, and 90 for the four classes, estimated coefficients of variation over 25 replications were approximately 30 percent higher on the average for misclassified matches and about the same for misclassified nonmatches.

The fact that the coefficients of variation decrease substantially as sample sizes increase indicates that calibration samples should be as large as possible. As the total number of records considered in these analyses was quite small, taking substantially larger samples was not practicable.

Examination of Table 6 shows that the estimated coefficients of variation associated with the cutoff weights using the modified information-theoretic procedure range from 15.3 percent to 99.5 percent; and from 14.3 percent to 115.4 percent with the basic information-theoretic procedure. The cutoff weights are consistent with the cutoff weights given in Table 4 and Table 5. Results in Tables 4 and 5 were obtained using the entire data set instead of samples.

Examination of Tables 7 and 8 show that the misclassification and nonclassification rates can vary significantly. Coefficients of variation of the estimated misclassification rates for the modified information-theoretic procedure vary from 33.2 to 109.9; for the basic procedure from 33.8 to 112.9.

Table 6: Estimated Cutoff Weights and Their Variances
25 Replications, With and Without Conditioning

Class	Status 1/	Estimated Cutoff Weights		Variance of Estimated Cutoff Weights		CVs of Estimated Cutoff Weights	
		LOWER	UPPER	LOWER	UPPER	LOWER	UPPER
1	C	2.66	7.72	7.02	2.05	99.5	18.5
2	C	1.44	1.44	0.62	0.62	54.9	54.9
3	C	-3.39	5.82	8.74	2.08	87.2	24.8
4	C	6.89	1.92	1.11	7.57	15.3	23.1
1	WC	-1.92	8.05	4.90	1.50	115.4	15.2
2	WC	-5.04	4.56	0.52	1.41	14.3	26.1
3	WC	-6.38	6.82	1.46	1.66	18.9	18.9
4	WC	1.71	12.13	3.11	7.56	102.9	22.7

1/ C-Conditioning, WC-Without Conditioning.

Table 7: Estimated Counts and Rates of Misclassification and Nonclassification
25 Replications, With and Without Conditioning

Class	Status 1/	Total Records	Misclassed as		Not Classed	Correctly Classed as		Proportion Misclassed as	
			Match	Non-Match		Match	Non-Match	Match	Non-Match
1	C	1021	10.4	27.4	75.2	260.7	647.2	.038	.041
2	C	624	9.7	3.0	0.0	244.0	367.3	.038	.008
3	C	256	3.0	3.5	94.2	85.2	70.0	.034	.048
4	C	344	1.4	10.2	23.5	54.3	254.6	.026	.039
Total		2245	24.5	44.1	192.9	644.2	1338.1	.037	.032
1	WC	1021	8.9	26.2	145.4	237.1	603.3	.036	.042
2	WC	624	3.8	3.9	450.6	89.4	76.3	.040	.048
3	WC	256	1.6	2.3	178.8	38.1	35.1	.041	.062
4	WC	344	1.3	9.6	57.7	38.8	236.6	.032	039
Total		2245	15.6	42.0	832.5	403.4	951.3	.037	.042

1/ C-Conditioning, WC-Without Conditioning.

Comparison of the modified and basic weighting procedures shows that the modified procedure is able to classify accurately significantly more records, particularly in classes 2 and 4, than the basic procedure. The results are consistent with those presented in Tables 4 and 5.

Results obtained using Efron's bootstrap imputation with 25, 100, 200, and 500 replications are consistent with the results in Tables 6, 7 and 8.

3.2.7. Overall Rate of Duplication

The overall rate of duplication (section 2.3.2) is 0.19 percent (100*102/(54850+102)) where the number of headquarters records is 54,850 and an estimated upper bound on the number of erroneous nonmatches is 102).

The estimated upper bound, 102, on the number of erroneous nonmatches is the number of matches

Table 8: Coefficients of Variation of Estimated Counts of Misclassification and Nonclassification 1/

25 Replications With and Without Conditioning

Class	Status 2/	Total Records	Misclassed as		Not Classed
			Match	Non-Match	
1	C	1021	69.5	47.4	54.7
2	C	624	64.6	81.1	0.0
3	C	256	96.6	84.1	40.9
4	C	344	109.9	33.2	60.8
1	WC	1021	62.3	42.3	34.0
2	WC	624	112.9	96.2	9.0
3	WC	256	106.9	65.5	8.1
4	WC	344	99.6	33.8	34.3

1/ Units are percentages.
2/ C-Conditioning, WC-Without Conditioning.

that are unblocked plus an upper bound on the the number that are erroneously classified as nonmatches during the discrimination stage. Thirty-nine records (section 3.1.2) are unblocked using the preferred set of blocking criteria.

The estimated upper bound consists of the sum of the estimated upper bounds on the numbers of automatically erroneously matched records in classes 1-4 and an estimate of the number of matches that are misclassified during manual review. The upper bounds at the 95 percent confidence level in classes 1-4 (using the estimates in Tables 7 and 8) are 24.9, 22.2, 8.9, and 4.5, respectively.

We assume that two percent of the estimated 124.3 matches in the estimated set of 192.9 records (see Tables 7 and 8) will be misclassed during manual review. This yields that 2.5 matches will be misclassed as nonmatches.

Thus, the upper bound is 102 (=39+24.9+22.2+8.9+4.5+2.5).

3.2.8. The Independence Assumption

Independence of comparisons does not hold. This is shown by the significant variation of the lower and upper cutoff weights across Classes 1 thru 4 in Tables 4, 5 and 6. If the comparisons were independent, then individual weights and cutoffs for the total weights would be reasonably consistent across classes. Individual weights (not shown) vary more than the cutoff weights across classes.

Independence of interactions within classes is illustrated by Tables 9 and 10. They show the two-way independence of the interactions of some of the subfields given in section 2.1.3 for subfields that are generally not connected and somewhat connected with blocking characteristics respectively. The variables used in the comparisons were defined in sections 2.1.3 and 2.2.2, respectively.

Table 9: Independence of Two-Way Interactions for Selected Subfields that are Generally Not Connected with Blocking Characteristics, By Class 1/

Class	K11/H	K22/H	K11/SN	K22/SN
1	yes	yes	no 2/	no 2/
2	NA	NA	yes	yes
3	no 4/	no 3/	no 2/	yes
4	yes	yes	yes	yes

NA- not applicable because one of two variables is basically the same as a blocking characteristic due to small sample size.

1/ Kii is the comparison of KEYWORDi with KEYWORDi, for i=1, 2; H is comparison of HOUSE NUMBER with HOUSE NUMBER; and SN is the comparison of STREET NAME with STREET NAME.

2/ Independent when H is included in a 3-way contingency table analysis.

3/ Independent when K11 is included.

4/ Independent when K22 is included.

Table 10: Independence of Two-Way Interactions for Selected Subfields that are Somewhat Connected with Blocking Characteristics, By Class

Class	W1/S1	W1/S2	W1/S3	W2/S1	W2/S2	W2/S3	W3/S1	W3/S2	W3/S3
1	yes	yes	yes	yes	yes	yes	yes	yes	yes
2	NA	yes	yes	NA	yes	yes	NA	yes	yes
3	no 1/	no 2/	no 3/	no 4/	no 2/	no 1/	no 5/	no 2/	no 1/
4	NA	NA	NA	yes	yes	no 1/	no 1/	no 2/	no 1/
A 6/	no	no	yes	yes	yes	yes	yes	yes	yes

NA- not applicable because one of two variables is used as a blocking characteristic.

1/ Independent when S2 is included in a 3-way contingency table analysis.

2/ Independent when S1 is included.

3/ Independent when W2 is included.

4/ Independent when W3 is included.

5/ Independent when S3 is included.

6/ Aggregate of Classes 1-4.

The Fellegi-Sunter model (1969, pp. 1189-1190) does not require full independence of interactions. It only requires that interactions be conditionally independent.

In over half the entries in Tables 9 and 10, the two-way interactions are independent unconditionally at the 95 percent confidence level and the hierarchical principle (Bishop, Fienberg, and Holland, 1975) assures that all such two-way interactions are always conditionally independent. In all cases in which two-way interactions are not unconditionally independent, a third variable was found so that the two-way interactions were independent at the 95 percent confidence level given the third variable.

It is important to note two points. First, some of the interaction of variables (not presented in the tables) such as H and S1 or W1 and K11 are often not independent unconditionally and it seems likely that they will generally not be independent conditionally. Second, building a precise model, by mutually exclusive class, in which only the minimal set of variables necessary for effective discrimination is included, and which precisely models the conditional relationships, is likely to be difficult and heavily dependent on the empirical data base used.

What we attempted to do in our approach was to find a superset of the minimal set of variables needed for effective discrimination; apply them all in creating the weights for each class; perform minimal modification in the basic procedures for creating the weights; and show that the failure of the independence assumption is not too crucial.

4. CONCLUSIONS AND FUTURE WORK

This section contains a brief summary of the results of this paper, a discussion of how the results relate to previous applied work and existing theory, and a set of problems for future research.

4.1. Summary

The results of this paper imply that the keys to delineating matches and nonmatches accurately are: (1) good spelling standardization and (2) accurate identification of corresponding subfields. They also imply that the independence assumption, required by the information-theoretic model of Fellegi and Sunter (1969), is not critical in practical applications of the type performed in this paper.

A key advantage of the Fellegi-Sunter approach is that it lends itself to incremental improvements, as knowledge of both file properties and data manipulation techniques (via software) increase.

4.2. Further Discussion of Results

4.2.1. Independent Application of Multiple Blocking Criteria

Newcombe et al. (1962, pp. 563-564) provide an example of applying multiple blocking criteria independently. They blocked first on surname and then on maiden name in files of individuals used for epidemiological research. In their study of a special sample of 3560 matches (linkages in their terminology), 98.4 percent (3504) were obtained using SOUNDEX coding of surname and an additional 1.4 percent (to a total 99.8 percent) were obtained using SOUNDEX coding of maiden surname. The increase in the total number of pairs considered for review when the second blocking criterion was used was 100 percent.

The results of section 3.1 show that, within the set of criteria considered, no single blocking criterion can yield a subset of pairs containing 80 percent of matches and no two can yield subsets containing 90 percent. The work of Winkler (1984a,b) provides a considerably more exhaustive study of blocking criteria and shows how the set of criteria used in this study work reasonably well on two additional sets of files.

Kelley (1985) provides a theoretical foundation for the simultaneous consideration of several subfields which is consistent with the Fellegi-Sunter model. In hypothetical examples, he shows how best to apply simultaneously first name, surname, and sex as blocking criteria. Section 3.1 results show that criterion 5, 10 characters of the NAME, does not perform well (62.4 percent of matches are not blocked and only 14.4 percent of the blocked pairs are matches) while criterion 1, 3 digits of the ZIP and 4 characters of the NAME, performs considerably better (45.5 percent of matches unblocked and 66.8 percent of the blocked pairs are matches). Thus, our results serve as partial corroboration of Kelley's results.

It seems likely that independent application of multiple blocking criteria such as done in this paper will be necessary to identify matches in other files of businesses. This is primarily due to lack of identifiers such as surnames.

4.2.2. Spelling Standardization

The comparison of Figures 1 and 2 in section 3.2.2 showed that improved spelling standardization of commonly occurring words did not yield any dramatic improvement in the ability to distinguish matches and nonmatches. Results for other classes (not shown) were similar. The results, however, may not be representative because the files had already been standardized using a somewhat more elementary set of tables. It is possible that improvements could be more dramatic when results using totally unstandardized files are compared with results using well standardized files.

Additionally, consistent spelling of commonly occurring words can allow their identification; thus, making it easier to identify other subfields having greater distinguishing power.

4.2.3. Subfield Identification

Section 3.2 results (particularly Figures 2-4) showed improvements in the Fellegi-Sunter weighting procedure's ability to delineate accurately matches and nonmatches and reduce the size of the manual review region. The improvements were due to the identification of subfields in the NAME and STREET fields using ZIPSTAN and KEYWORD software, respectively.

The improvements using ZIPSTAN in classes 1 and 4 (not shown) were quite substantial. They were, however, not as dramatic as the improvements in classes 2 and 3 when conditioning procedures were used.

The results basically show us that it may be possible to delineate and compare subfields (particularly within the NAME field) that yield greater distinguishing power. In particular, if such comparable subfields are distinguished, then string comparator metrics (see e.g., Winkler, 1985) which allow assignment of weights of partial agreement between strings (rather than just 1-agree and 0-disagree) could be used to deal with subfields containing minor keypunch/transcription errors.

4.2.4. Independence, Conditioning, and Steepest Ascent

The results in section 3.2 (particularly subsections 3.2.1 and 3.2.8) show that the comparisons of characteristics of various subfields are generally not independent. Fellegi and Sunter (1969, p. 1191) indicate that their weighting scheme may work well in practice even when the independence assumption is not met.

In an early analysis (not shown), weights were computed uniformly over all pairs within the set of blocked pairs, rather than separately in the four subclasses. Analyses similar to those in section 3.2 (particularly, using figures like Figures 1-5) showed that weights computed uniformly did not have as much distinguishing power. In particular, the curves of nonmatches and matches never moved as far apart as the curves moved apart in Figure 5. Results (not shown) for other classes used in this paper were quite similar to those in Figures 1-5.

We can conclude that, at least in our example, dependence of comparisons leads to less discriminating power. We should note, however, that a large number of comparisons were performed, some of which are likely not to be independent conditionally. It may be possible that subsets of the comparisons (they are likely to vary significantly from class to class) may be created in which the comparisons are conditionally independent. For such subsets, however, it is not clear whether the overall discriminating power will increase.

It is important to note that, for those procedures in which only one blocking criterion is used (such as blocking on SOUNDEX abbreviation of surname in files of individuals), it may be possible to compute weights uniformly over the entire set of blocked pairs. The four classes which we considered were created using the preferred set of four blocking criteria. Thus, our weight creation scheme is conditional on the set of blocking criteria.

The conditioning arguments in this paper consisted primarily of the subdivision of the set of blocked pairs into four classes based on the four blocking criteria and steepest ascent methods of weight variation. Both procedures are cumbersome to apply, the second particularly so. It may be possible to produce some algorithm for conditioning or some other method which allows a systematic approach to conditioning. Bishop, Fienberg, and Holland (1975, Chapter 11) provide a useful discussion of the difficulties with some of the measures of association that have been developed.

4.2.5. Legitimate Representation Differences and Keypunch/Transcription Error

Fellegi and Sunter (1969, pp. 1193-1194) provided a specific model which incorporates error rates associated with legitimate representation differences of the same entity (see e.g., the name variations in section 2.1.3) and/or keypunch/transcription error. Their results (see also Coulter, 1977; Kirkendall, 1985) show that, in the presence of such errors, agreement weights remain approximately the same as agreement weights in the absence of such errors, while disagreement weights (which are generally negative) increase. The results have substantial intuitive appeal.

Review of figures like Figures 1-5 for classes 1, 3, and 4 (not shown) and examination of pairs that are either misclassified or not classified in all 4 classes indicate that keypunch error plays a substantially greater role in classes 1 and 3 than in classes 2 and 4. The results are consistent with Table 4 results in which all records in classes 2 and 4 are classified (none held for manual review) while a moderate number of records in classes 1 and 3 (55 of 1021 and 42 of 256, respectively) are held for manual review.

A partial explanation of the differences is that classes 1 and 3 contain a moderate number of pairs of records having substantial variations in the NAME and/or STREET fields while classes 2 and 4 do not. In class 1, many keypunch errors occur after the first four characters of the NAME. Being able to block on TELEPHONE (class 3), allows significant reduction in the number of erroneous nonmatched because so many keypunch/transcriptions can occur in the NAME and STREET fields (see also Winkler, 1984a).

An additional series of steepest ascent variations were performed in classes 1 and 3. In all cases, the distinguishing power remained constant or became slightly worse. In some cases, graphs such as given by Figure 5 contained curves of nonmatches and matches for which the humps moved apart but for which the manual review region remained constant or increased in height. Thus, it seems unlikely that more conditioning in the form presented in this paper will improve procedures. Rather, it seems likely that improvements will depend more on better identification and comparison of subfields.

4.2.6. Adaptability of the Fellegi-Sunter Procedures

Newcombe et al. (1959, 1962) first showed that the basic weighting procedure as presented in Fellegi and Sunter (1969) could be improved by adapting it to make use of additional comparative information. Figures 1-5 in this paper illustrate successive improvements which can be obtained using spelling standardization, additional comparisons of subfields of the NAME and STREET fields, and conditioning arguments.

Further improvements seem likely. They can be obtained using techniques that are already available. For instance, Statistics Canada (1982) has developed sophisticated methods of delineating subfields within the NAME field for use on the Canadian Business Register. Identifying subfields as Statistics Canada has done could allow a number of less sophisticated comparisons (such as first four characters and next six characters of the NAME field) to be dropped and discriminating power to increase. ZIPSTAN software (U.S. Dept. of Commerce, 1978b) yielded subfields of the STREET field which provided increased discriminating power.

Use of frequency counts of the occurrence of substrings (e.g., Zabrinsky occurs less often and has more distinguishing power than Smith) could be incorporated in matching lists of businesses. Presently, such matching using frequency counts is applied to lists of individuals (e.g., U.S. Dept. of Agriculture, 1979; U.S. Dept. of Commerce, 1978a). The theoretical justification for procedures using frequency-based matching are explicitly described by Fellegi and Sunter (1969, pp. 1193-1194).

Use of frequency-based matching involves use of lookup tables for obtaining weights associated with individual comparisons. Such lookups can be performed efficiently using K-D trees (Friedman, Bentley, and Finkel, 1977). EIA presently uses K-D trees for search of lookup tables during spelling standardization.

String comparator metrics (see e.g., Winkler, 1985) allowing comparison of strings containing minor keypunch errors could also be used in adapting the weighting procedures.

4.3. Problems Remaining

Effective evaluation of the efficacy of various matching procedures requires having a representative data base in which matches and nonmatches have been identified and tracked. Such data bases can be created during list updating projects and are necessary if incremental improvements in procedures are to be made (see e.g., Coulter and Mergerson, 1977; Smith et al., 1983).

Effective evaluation also requires having common terminology and measures that allow rough comparison of results obtained using significantly different data bases and/or methodologies. The results of this paper and others (see e.g., Newcombe et al., 1983; Rogot et al., 1983) suggest a number of avenues for future research that can be incorporated into existing procedures in a straightforward manner.

4.3.1. Error Rates

Various authors (see e.g., Newcombe et al., 1983; Rogot et al., 1983) have presented the rates of erroneous matches and nonmatches during the discrimination stage but generally do not mention the rates of erroneous nonmatches that remain unlinked during the blocking stage. As the Fellegi-Sunter model explicitly provides measures of the Type I and Type II error rates, it seems natural to extend investigation of such rates to both blocking and discrimination stages.

The results of this paper imply that error rates occurring during both stages must be investigated simultaneously. For instance, during early stages of the work at EIA no effective methods existed for accurately delineating matches and nonmatches during the discrimination stage. As more effective methods of delineating matches and nonmatches during the discrimination stage are developed, it seems likely that additional blocking criteria (such as criterion 5 in section 3.1) may be adopted without increasing the rate of erroneous nonmatches.

Other measures, such as the overall rate of duplication given in this paper (see also Winkler, 1984a,b), may provide additional insight into how well a specific application is performed and provide additional information comparable with other applications.

Type I error rates based on samples (see e.g., Winkler, 1984a,b) have been shown to yield coefficients of variations of approximately 100 percent even with samples as large as 1800. Although Fellegi and Sunter (1969) indicate that estimating error rates based on samples yields high variances, they did not provide an example showing the magnitude of the problem. There may be better methods for obtaining such error rates and their variances when samples are used.

4.3.2. General Applicability of Linkage Mechanisms

Winkler (1984a,b) showed that the preferred set of blocking criteria are reasonably applicable to two other data bases having different characteristics from the empirical data base that was used for analyses in this paper. In those papers, however, blocking criteria were investigated independent of the discrimination stage.

Investigations of the efficacy of different blocking strategies when both blocking and discrimination stages are considered simultaneously are necessary. The investigations should be performed on files with significantly different characteristics.

For instance, is the use of an abbreviation method such as SOUNDEX (e.g., Bourne and Ford, 1961) or NYSIIS (e.g., Lynch and Arends, 1977) abbreviation of SURNAME the only way to block files of individuals? If so, why are such blocking methods effective in reducing the rate of erroneous nonmatches? What methods were investigated and why were they rejected? Should files of individuals be blocked several different ways using significantly different blocking criteria?

4.3.3. String Comparators

If corresponding strings such as SURNAME are identified, then it is possible to define distance or weighting functions that compare nonidentical strings. Such weighting functions (see e.g. Winkler, 1985, pp. 12-16) can be derived using abbreviation methods such as SOUNDEX (e.g., Bourne and Ford, 1961), using the Damerau-Levenstein metric (e.g., Hall and Dowling, 1980, pp. 388-390), or the string comparator of Jaro (e.g., U.S. Dept of Commerce, 1978a, pp. 83-101).

Each of the methods is intended to allow comparison of strings in which minor typographical differences occur. What are the relative merits of different weighting functions? Are there any better algorithms for string comparison?

4.3.4. Tracking True and False Matches

In linking pairs of records in lists of businesses, many erroneous matches will have similar NAMEs and/or STREET ADDRESSes. Matches may have different NAMEs and/or STREET ADDRESSes (e.g., subsidiaries, successors). Delineation of most such matches and nonmatches can require manual followup which is both time-consuming and expensive.

If matches and nonmatches are tracked properly and the weighting methodology for delineating matches and nonmatches is reasonably effective, then many nonmatches that have similar NAMES and STREET ADDRESSes to previous nonmatches or matches having different NAMES and/or STREET ADDRESSes from their true parents will not require manual review.

To determine if it is cost-effective to track matches and nonmatches, research is needed to show:

1. how classes of matches and nonmatches of records linked using various blocking criteria should be set up to allow tracking;
2. how effective weighting schemes should be determined that allow maximum use of the tracking system;
3. how pairs newly linked during an update should be compared within equivalence classes and across equivalence (a record can be linked truly once and falsely many times);
4. how updating using the results of 1, 2, and 3 should be performed; and
5. how the results of the updating should be evaluated.

ACKNOWLEDGMENT

The author wishes to thank Dr. Yvonne M. Bishop for a number of helpful comments and conversations.

REFERENCES

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), Discrete Multivariate Analysis, MIT Press, Cambridge, MA.

Bourne, C. P., and Ford, D. F. (1961), "A Study of Methods for Systematically Abbreviating English Words and Names," J. ACM, 8, 538-552.

Coulter, R.W. (1977), "An Application of a Theory for Record Linkage," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Coulter, R.W. and Mergerson, J.W. (1977), "An Application of a Record Linkage Theory in Constructing a List Sampling Frame," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Cochran, W.G. and Cox, G.M. (1957) Experimental Designs, J. Wiley and Sons, New York.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," Ann. Stat., 7, 1-26.

Efron, B. and Gong, G. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," The American Statistician, 37, 36-48.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," JASA, 40, 1183-1210.

Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977), "An Algorithm for Finding Best Matches in Logarithmic Expected Time," ACM Trans. Math. Software, 3, 209-226.

Hall, P. A. V. and Dowling, G. R. (1980), "Approximate String Matching," Computing Surveys, 12, 381-402.

Herzog, T. and Rubin, D. (1983), "Using Multiple Imputations to Handle Nonresponse," in Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies, edited by Madow, W.G., Olkin, I., and Rubin, D.B. Academic Press, New York, 210-245.

Kelley, R. P. (1985), "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," Invited paper presented at the Workshop on Exact Matching Methodologies in Rosslyn, VA, on May 9-10, 1985.

Kirkendall, N. (1985). "Weights in Computer Matching: Applications and an Information Theoretic Point of View," Record Linkage Techniques--1985, Internal Revenue Service.

Lynch, B.T. and Arends, W.L. (1977), "Selection of a Surname Coding Procedure for the SRS Record Linkage System," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959), "Automatic Linkage of Vital Records," Science, 130, 954-959.

Newcombe, H.B. and Kennedy, J.M. (1962), "Record Linkage," Communications of the ACM, 5, 563-566.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J., Strugnell, A., and Abbatt, J.D. (1983), "Reliability of Computerized Versus Manual Searches in a Study of the Health of Eldorado Uranium Workers," Comput. Biol. Med., 13, 157-169.

Pollock, J. and Zamora, A. (1984), "Automatic Spelling Correction in Scientific and Scholarly Text," Communications of the ACM, 27, 358-368.

Rubin, D. (1978), "Multiple Imputations in Sample Surveys--A Phenomenological Bayesian Approach to Nonresponse," ASA 1978 Proceedings of the Section on Survey Research Methods, 20-28.

Rogot, E., Schwartz, S., O'Conor, K., and Olsen, C. (1983), "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index." ASA 1983 Proceedings of the Section on Survey Research Methods, 319-324.

Smith, M., Newcombe, H.B., and Dewar, R. (1983), "Automated Nationwide Death Clearance of Provincial Cancer Registry Files--The Alberta Cancer Registry Study," ASA 1983 Proceedings of the Section on Survey Research Methods, 300-305.

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

U. S. Department of Agriculture/ Statistical Reporting Service (1979), "List Frame Development: Procedures and Software."

U. S. Department of Commerce, Bureau of the Census/Agriculture Division (1981), "Record Linkage for Development of the 1978 Census of Agriculture Mailing List."

U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978a), "UNIMATCH: A Record Linkage System."

U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978b), "ZIPSTAN: Generalized Address Standardizer."

Winkler, W. E. (1984a), "Issues in Developing Frame Matching Procedures: Exact Matching Using Elementary Techniques." Presented to the ASA Energy Statistics Committee in April 1984.

Winkler, W. E. (1984b), "Exact Matching Using Elementary Techniques." ASA 1984 Proceedings of the Section on Survey Research Methods, 237-242.

Winkler, W. E. (1985), "Preprocessing of Lists and String Comparison," Record Linkage Techniques--1985, Internal Revenue Service.