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		Street Name	Suffixes		Unit
		1	2		1	2	
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.	1ØØ	N		COURT	SQ		
8.	1ØØ	CT	SQ	*** NO NAME ***			RM 167
9.	2589			WILLIAMS	DR		AP 6
1Ø.	15			RAILROAD	AV		
11.				2ND	AV	HW	1Ø
12.				MAIN	ST		
13.	184	N		DU PONT	PW		
14.	123Ø			16TH	ST		
15.	48Ø			*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):

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

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.