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.	Pre- fixes 1	2	Street Name	Suf- fixes 1	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	W 31st St N Church St	
	Hardsburg	PA 18207	713/643-2121
	Smith J K	N Church St	
	Hardsburg	PA 18207	missing
(d)	Mcneely R	3312-14 Harris Ave	
	MPLS	MN 55246	612/929-6677
	R Mcden Neely	3312 Harris Ave	
	St Louis Par	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

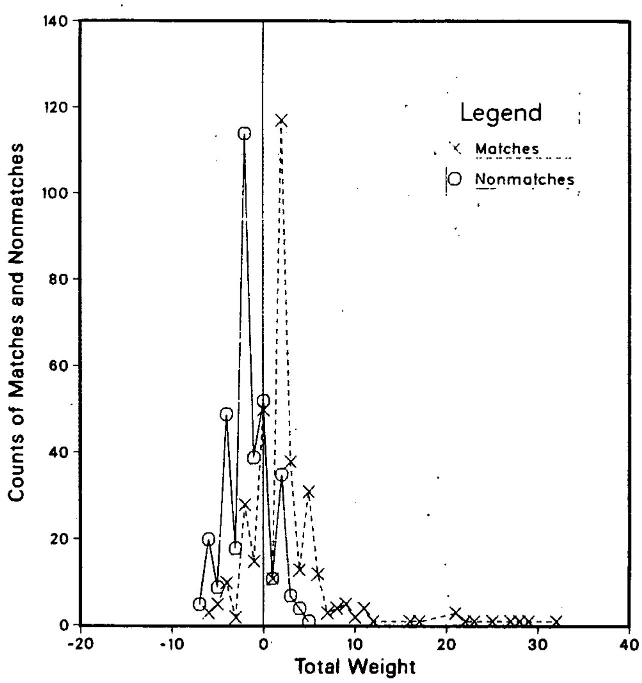
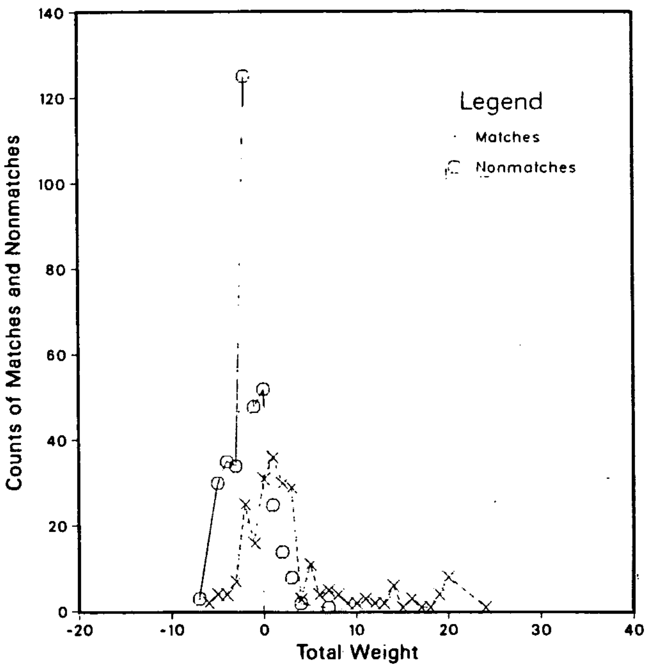


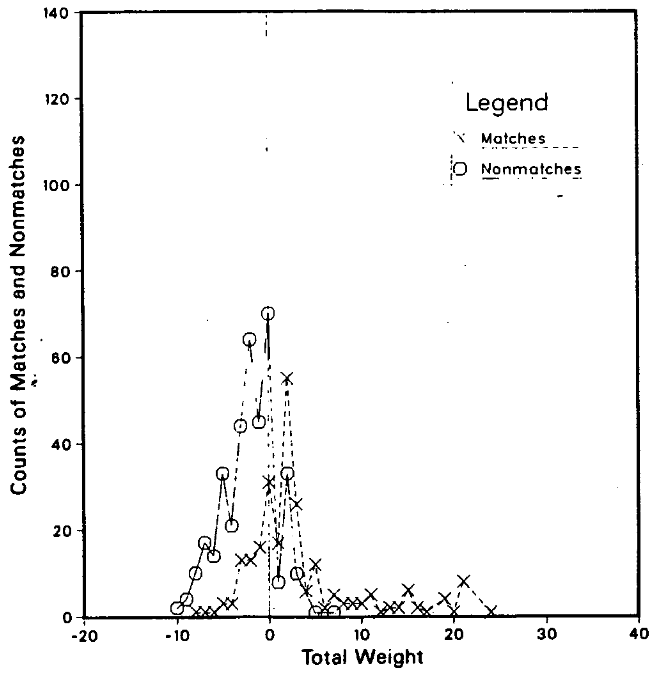
FIGURE 2: Total Weight Versus
Counts of Matches and Nonmatches
After New Spelling Standardization
Prior to Identification of Subfields



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.)

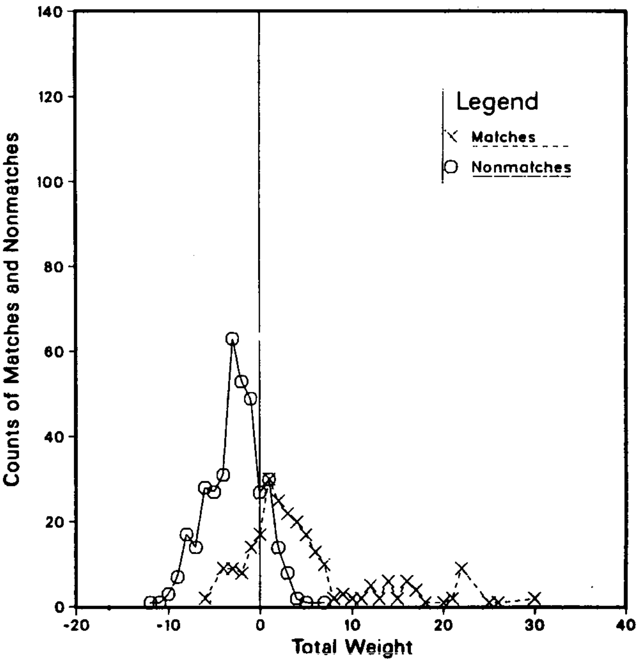
Although not shown in this paper, examination of similar sets of plots for other classes, particularly those blocked using the NAME field, show less improvement when additional weights obtained using the ADDRESS subfields are used.

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.

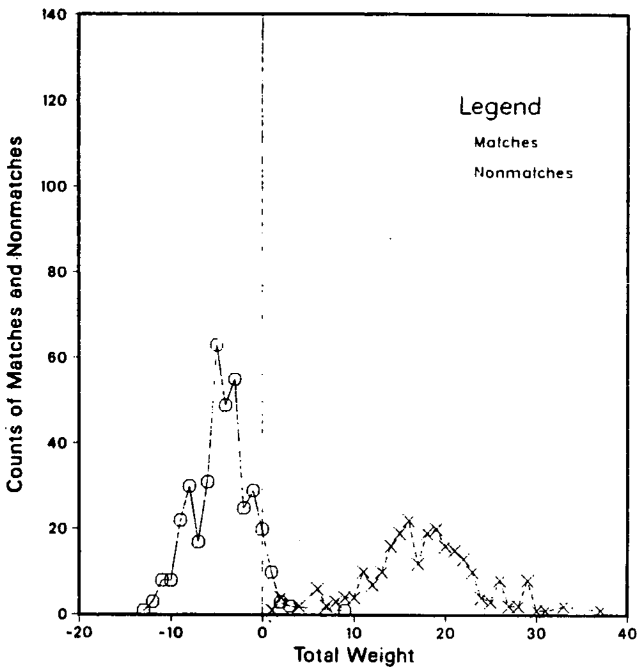
FIGURE 4: Total Weight Versus Counts of Matches and Nonmatches After New Spelling Standardization Name and Address Subfield Identification



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.