Survey Methodology, June 1989
Vol. 15, No. 1, pp. 101–117
Statistics Canada

Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage

WILLIAM E. WINKLER[1]

ABSTRACT

Let A × B be the product space of two sets A and B which is divided into **matches** (pairs representing the same entity) and **nonmatches** (pairs representing different entities). Linkage rules are those that divide A × B into **links** (designated matches), **possible links** (pairs for which we delay a decision), and **nonlinks** (designated nonmatches). Under fixed bounds on the error rates, Fellegi and Sunter (1969) provided a linkage rule that is optimal in the sense that it minimizes the set of possible links. The optimality is dependent on knowledge of certain probabilities that are used in a crucial likelihood ratio. In applying the record linkage model, an independence assumption is often made that allows estimation of the probabilities. If the assumption is not met, then a record linkage procedure using estimates computed under the assumption may not be optimal. This paper contains an examination of methods for adjusting linkage rules when the independence assumption is not valid. The presentation takes the form of an empirical analysis of lists of businesses for which the truth of matches is known. The number of possible links obtained using standard and adjusted computational procedures may be dependent on different samples. Bootstrap methods (Efron 1987) are used to examine the variation due to different samples.

KEY WORDS: Decision rule; Error rate; Steepest ascent; Bootstrap; Capture-recapture.

1. INTRODUCTION

This paper presents an analysis of decision rules obtained by applying the Fellegi-Sunter model of record linkage to lists of businesses. The analysis compares a rule obtained under an independence assumption that is typically assumed in practice with rules that include methods for adjusting for the failure of the independence assumption.

Given two lists, we wish to use identifying information to delineate those record pairs that represent the same entities (**matches**) and those that are different (**nonmatches**). Thus, we desire to define a linkage rule that allows us to divide the cross-product space of pairs into **links** (designated matches), **possible links** (pairs for which a decision is delayed), and **nonlinks** (designated nonmatches).

Under fixed bounds on the numbers of erroneous matches and nonmatches, Fellegi and Sunter (1969, Theorem) provide a procedure that, in theory, minimizes the number of possible links. The optimality is dependent on knowledge of certain probabilities that are used in a crucial likelihood ratio.

In typical applications, an independence assumption is made that allows estimation of the probabilities used in the likelihood ratio. The probabilities are called **matching parameters.** If the independence assumption is not valid (Winkler 1985c; Kelley 1986) then linkage rules based on the estimated probabilities may not be optimal.

[1] William E. Winkler, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233, USA.

Given fixed bounds on error rates, **better** linkage rules will be those that reduce the set of possible links. If a rule is based on matching parameters that are estimated under an invalid independence assumption, then it may be possible to develop adjustment procedures to determine better rules. To test whether one rule is statistically better than another, we use Efron's bootstrap (1987; also Hall 1988).

The remainder of the paper presents background, methods, and results from applying several record linkage rules to lists of businesses. The application involves pairs of lists for which the truth and falsehood of linkages are known.

The second section of this paper is divided into four subsections. The first contains a description of the data base and the specific subfields that are compared. The second subsection contains a summary of the Fellegi-Sunter model. The third subsection highlights common assumptions made and computational procedures used. It also contains details of computational procedures that are specific to the application of this paper.

The fourth subsection describes the evaluation procedures. The basic evaluation technique involves comparing sizes of the regions of possible links when different types of linkage rules are applied under fixed error bounds. The sizes of the regions of possible links are statistics that may be dependent on the samples used in calibrating the linkage rules. Efron's bootstrap (1987, 1982, 1979; also Hall 1988) is used to evaluate their distributions.

Results are presented in the third section. This is followed in the fourth section by discussion of the robustness of weight adjustment procedures, the type of conditioning represented by the adjusted weights, additional types of comparisons, and the use of extra blocking criteria. Finally, the paper concludes with a summary.

2. DATA BASE, LINKAGE MODEL, COMPUTATIONAL AND EVALUATION PROCEDURES

2.1 Data Base

The description of the data base is divided into two components. The first component is a description of the overall properties. The second contains a listing of the specific subfield comparisons that are made.

2.1.1 Overall Description

The data base of 57,900 records contains 54,850 records that are identified as individual companies and 3,050 duplicates. A pair of records that consists of a company and its corresponding duplicate is a match; all others are nonmatches.

The data base was constructed from 11 Energy Information Administration (EIA) and 47 State and industry lists containing 176,000 records. Duplicates were identified via elementary techniques, through call-backs (phone numbers are sometimes present) and through surveying.

The decision rules that are developed are only applied to those pairs that generally represent hard-to-identify duplicates. Easy-to-identify duplicates are those pairs having substantial portions of their name and addresses agreeing on a character-by-character basis.

An example of a hard-to-identify duplicate might be:

NAME	STREET	CITY	STATE	ZIP
Zabrinsky Fuel	16 W Sycamore St	Dayton	OH	53315
Zabrinky Cmpny	167 Sycamere St	Springfield	OH	53315.

We observe that both 'Zabrinsky' and 'Sycamore' are spelled wrong in the second record, that 'Cmpny' is a nonstandard abbreviation, and that Springfield OH, a suburb of Dayton, has Postal ZIP code 53315.

2.1.2 Specific Subfields Compared

There are four sets of specific subfields that are compared in each pair of records. First are those that can be obtained through easy substring comparisons. For instance, we could compare character positions 1-4 of the NAME field from one record with the corresponding same character positions of the NAME field in another record.

In Table 1 WL-NAME is obtained by sorting the NAME field by words of decreasing length with ties broken by an alpha sort. Corresponding subfields are then compared on a character-by-character basis.

The second set is the four comparisons of the first and second largest words in the NAME field. Ties are again broken by an alpha sort.

The last two sets are of subsets of the STREET and NAME fields that are designated by highly sophisticated software. ZIPSTAN software from the Census Bureau (U.S. Dept. of Commerce 1978b) is used to obtain corresponding subfields of the STREET field. The subfields are: House No., Prefixes 1 and 2, Street Name, Suffixes 1 and 2, and Unit. Prefixes are directions such as East and North. Suffixes are words such as Street and Road. Unit designates identifiers such as apartment or suite number.

The NSKGEN5 module from software used in the Canadian Business Register (Statistics Canada 1984, 1982) is used to obtain corresponding subfields of the NAME field. NSKGEN5 creates three groups of words. The first group consists of three abbreviations with the first corresponding to surname if present. The second group contains two words with the first corresponding to surname. The third group is a single word obtained by concatenating and abbreviating individual words in the NAME field. Details are given in Winkler (1987) or in Statistics Canada (1984, 1982).

2.2 Fellegi-Sunter Model

The Fellegi-Sunter Model uses a decision-theoretic approach establishing the validity of principles first used in practice by Newcombe (Newcombe *et al.* 1959). To give an overview, we describe the model in terms of ordered pairs in a product space. The description closely follows Fellegi and Sunter (1969, pp. 1184-1187).

Table 1

Corresponding Subfields Compared on a Character-by-Character Basis

Field	1-4, 5-10, 11-20, 21-30
NAME	1-4, 5-10, 11-20, 21-30
STREET	1-6, 7-15, 16-30
ZIP	1-3, 4-5
CITY	1-5, 6-10, 11-15
STATE	1-2
TELEPHONE	1-3, 4-6, 7-10
WL-NAME	1-4, 5-10, 11-20, 21-30

There are two populations A and B whose elements will be denoted by a and b. We assume that some elements are common to A and B. Consequently the set of ordered pairs

$$A \times B = \{(a,b): a \in A,\ b \in B\}$$

is the union of two disjoint sets of **matches**

$$M = \{(a,b): a = b,\ a \in A,\ b \in B\}$$

and **nonmatches**

$$U = \{(a,b): a \neq b,\ a \in A,\ b \in B\}.$$

The records corresponding to members of A and B are denoted by $\alpha(a)$ and $\beta(b)$, respectively. The **comparison vector** γ associated with the records is defined by:

$$\gamma[\alpha(a), \beta(b)] \equiv \{\gamma^1[\alpha(a), \beta(b)], \gamma^2[\alpha(a), \beta(b)], \ldots, \gamma^K[\alpha(a), \beta(b)]\}.$$

Each of the γ^i, $i = 1, \ldots, K$, represents a specific comparison. For instance, γ^1 could represent agreement/disagreement on sex. γ^2 could represent the comparison that two surnames agree and take a specific value or that they disagree.

Where confusion does not arise, the function γ on A × B will be denoted by $\gamma(\alpha,\beta)$, $\gamma(a,b)$, or γ. The set of all possible realizations of γ is denoted by Γ.

The conditional probability of $\gamma(a,b)$ if $(a,b) \in M$ is given by

$$m(\gamma) \equiv P\{\gamma[\alpha(a)\beta(b)] \mid (a,b) \in M\}$$

$$= \sum_{(a,b) \in M} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a,b) \mid M].$$

Similarly we denote the conditional probability of γ if $(a,b) \in U$ by $u(\gamma)$.

We observe a vector of information $\gamma(a,b)$ associated with pair (a,b) and wish to designate a pair as a link (denote the decision by A_1), a possible link (decision A_2), or a nonlink (decision A_3). A **linkage rule** L is defined a mapping from Γ, the comparison space, onto a set of random decision functions $D = \{d(\gamma)\}$ where

$$d(\gamma) = \{P(A_1 \mid \gamma), P(A_2 \mid \gamma), P(A_3 \mid \gamma)\};\ \gamma \in \Gamma$$

and

$$\sum_{i=1}^{3} P(A_i \mid \gamma) = 1.$$

There are two types of error associated with a linkage rule. A **Type I error** occurs if an unmatched comparison is erroneously linked. It has probability

$$P(A_1 \mid U) = \sum_{\gamma \in \Gamma} u(\gamma) \cdot P(A_1 \mid \gamma)$$

A **Type II error** occurs if a matched comparison is erroneously not linked. It has probability

$$P(A_3|U) = \sum_{\gamma\epsilon\Gamma} m(\gamma) \cdot P(A_3|\gamma).$$

Fellegi and Sunter (1969) define a linkage rule L_0, with associated decisions A_1, A_2, and A_3, that is optimal in the following sense:

THEOREM (Fellegi-Sunter 1969). Let L' be a linkage rule with associated decisions A_1', A_2', and A_3' such that it has the same error probabilities $P(A_3'|M) = P(A_3|M)$ and $P(A_1'|U) = P(A_1|U)$ as L_0. Then L_0 is optimal in that $P(A_2|U) \leq P(A_2'|U)$ and $P(A_2|M) \leq P(A_2'|M)$.

In other words, if L' is any competitor of L_0 having the same Type I and Type II error rates (which are both conditional probabilities), then the conditional probabilities (either on set U or M) of not making a decision under rule L' are always greater than under L_0. L_0 is described in subsection 2.3.1.

The Fellegi-Sunter linkage rule is actually optimal with respect to any set Q of ordered pairs in $A \times B$ if we define error probabilities P_Q and a linkage rule L_Q conditional on Q. Thus, it may be possible to define subsets of $A \times B$ on which we make use of differing amounts and types of available information.

For instance, if we have a set of pairs in which telephone number is present, we might use telephone number and a few characters from the name to designate links. With other pairs, we may additionally have to utilize information from the street address and the city name.

Sets of ordered pairs Q on which the Fellegi-Sunter linkage rule is applied are often obtained by **blocking criteria**. Blocking criteria are sort keys that are used to reduce the number of pairs that are considered. Rather than consider all pairs in $A \times B$, we might only consider pairs that agree on the first three digits of the ZIP code or on a suitable abbreviation of surname.

2.3 Computational Procedures

This section is divided into five parts. The first part contains a description of the general linkage rule of the Fellegi-Sunter Model. The second contains a description of the simplified computational procedures when a conditional independence assumption is made.

Background on the validity of the conditional independence assumption is presented in the third part. The fourth describes two general methods of adapting computational procedures. The fifth provides a description of the specific computational procedures of this paper.

2.3.1 General Form of Linkage Rule

To provide a background for understanding why specific computational procedures are used, we consider the following likelihood ratio

$$R \equiv R[\gamma(a,b)] = m(\gamma)/u(\gamma). \tag{2.1}$$

We observe that, if γ represents a comparison of K fields, then there are at least 2^K probabilities of form $m(\gamma)$. If γ represents agreements of K fields, we would expect this to occur more often for matches M than for nonmatches U. The ratio R would then be large. Alternatively, if γ consists of disagreements, the ratio R would be small.

If the numerator is positive and the denominator is zero in (2.1), we assign an arbitrary very large number to the ratio. The Fellegi-Sunter linkage rule takes the form:

$$\begin{aligned}&\text{If } R > \text{UPPER, then denote } (a,b) \text{ as a link.}\\ &\text{If LOWER} \leq R \leq \text{UPPER, then denote } (a,b) \text{ as a possible link.}\\ &\text{If } R < \text{LOWER, then denote } (a,b) \text{ as a nonlink.}\end{aligned} \tag{2.2}$$

The cutoffs LOWER and UPPER are determined by the desired error rate bounds.

2.3.2 Simplification Under Conditional Independence Assumption

In practice, computation is simplified two ways. The first is by the conditional independence assumption of Fellegi and Sunter (1969):
For each $\gamma \in \Gamma$

$$\begin{aligned}m(\gamma) &= m_1(\gamma^1) \cdot m_2(\gamma^2) \ldots m_K(\gamma^K) \text{ and}\\ u(\gamma) &= u_1(\gamma^1) \cdot u_2(\gamma^2) \ldots u_K(\gamma^K)\end{aligned}$$

where for $i = 1, 2, \ldots, K$

$$\begin{aligned}m_i(\gamma^i) &= P(\gamma^i | (a,b) \in M) \text{ and}\\ u_i(\gamma^i) &= P(\gamma^i | (a,b) \in U).\end{aligned}$$

This assumption basically is that agreement on one characteristic such as surname does not depend on agreement of other characteristics such as house number or age.

The second is to use a computationally convenient function of the ratio in (2.1). Log_2 is used. We then have

$$\begin{aligned}W \equiv W(\gamma) &= \text{Log}_2[m(\gamma)/u(\gamma)]\\ &= W^1 + W^2 + \ldots + W^K,\end{aligned} \tag{2.3}$$

where $W^i \equiv \text{Log}_2[m_i(\gamma^i)/u_i(\gamma^i)]$ for $i = 1, 2, \ldots, K$. We call W the **total comparison weight** associated with a pair and W^i, $i = 1, 2, \ldots K$, the **individual comparison weights.**

For the remainder of the paper we will assume that each component γ^i, $i = 1, 2, \ldots K$, in γ represents a two-state comparison (*e.g.*, agree/disagree). For convenience, we denote agreement in the *i*th component by γ^i_o, $i = 1, 2, \ldots K$. Under the conditional independence assumption, for each $i = 1, 2, \ldots K$, we need to estimate probabilities of the forms

$$P(\gamma = \gamma^i_o | M) \text{ and } P(\gamma = \gamma^i_o | U). \tag{2.4}$$

Using a set of pairs for which the truth and falsehood of matches are known, for each agreement γ^i_o, $i = 1, 2, \ldots, K$, we divide the set into the four subsets determined by the agree/disagree and match/nonmatch statuses in (2.4) to perform the estimation.

If no conditional independence assumption is made, we need to estimate $2 \cdot (2^K - 1)$ probabilities of form (2.1) and divide the set of pairs for which truth and falsehood are known to $2 \cdot (2^K - 1)$ subsets. Even with a small number of comparisons (say, 6 or less), we may not be able to obtain sufficiently large samples to allow accurate estimation of the probabilities.

2.3.3 Validity of Conditional Independence Assumption

Winkler (1985c) has shown that the independence assumption is not valid for simple comparisons of portions of the name and street address fields for list of businesses. Using similar portions of the name and street fields, Kelley (1986) has shown that the independence assumption is not valid for files of individuals. Furthermore, Kelley and Winkler have each shown that matching efficacy is sensitive to the set of pairs over which probabilities of the form (2.4) are computed.

Fellegi and Sunter indicate that, if the conditional independence assumption is not valid, then estimates of weights that are obtained via formula (2.3) will lose their strict probabilistic interpretation. By this, they mean that the linkage rule of their theorem may not actually minimize the number of possible links. They indicate that they believe their procedure to be robust to departures from the independence assumption.

Under the independence assumption, probabilities are computed as products of probabilities of the form (2.4). If we have a set of pairs for which truth and falsehood of matches are known, then we can adjust probabilities of form (2.4) for departures from the independence assumption. If the total weights obtained by adjustment yield substantially smaller sets of potential links under fixed bounds on error rates, then the Fellegi-Sunter procedure may not be robust to departures from independence.

2.3.4 General Adjustments

There are two general adjustments to the basic methods of computing individual comparison weights. The first consists of dividing the subset of pairs in $A \times B$ over which individual comparison weights are computed into several subsets. The linkage rule is obtained by restricting the basic Fellegi-Sunter rule to correspond to the different subsets on which weights are computed. Individual comparison weights may vary significantly in different subsets.

The second adjustment consists of modifying individual comparison weights. Under the independence assumption, we consider the equation

$$\begin{aligned} W &\equiv \mathrm{Log}_2(P(\gamma \in B_1 \cap B_2 \cap \ \ldots \ \cap B_K | M) / P(\gamma \in \mathrm{B}_1 \cap B_2 \cap \ \ldots \ \cap B_K | U)) \\ &= W^1 + W^2 + \ldots \ + W^K, \end{aligned}$$

where, for $i = 1, 2,$ and K, $W^i \equiv \mathrm{Log}_2(P(\gamma \in B_i | M) / P(\gamma \in B_i | U))$ and B^i is the set $\{\gamma^i = \gamma_0^i\}$ or its complement. We wish to find computationally tractable methods of adjusting the W^i, $i = 1, 2, \ldots, K$, so that their sum yields better linkage rules.

If there is a sample for which the truth and falsehood of matches are known, then we can estimate individual comparison weights (Tepping 1968) and the adjustments.

The simplest adjustment procedure involves a steepest ascent approach (*e.g.*, Cochran and Cox 1957). To begin, we use the known truth and falsehood of matches within a sample to estimate probabilities of the form (2.4). The probabilities are then used in computing individual comparison weights that are added to obtain an estimate of total weight (2.3). For each pair of fixed bounds on Type I and Type II errors, the cutoffs UPPER and LOWER of (2.2) can be determined. The number of potential links for rules of the form (2.2) follows immediately.

Next, we chose an individual comparison weight, change it by a fixed amount (say ± 1), recompute the total weight (2.3) using the new individual weight, and find new cutoffs UPPER and LOWER and a new region of potential links.

If under fixed bounds of errors, the size of the region of possible links decreases, then we continue adjusting the individual comparison weight (either up or down) until the region ceases its decrease in size. We continue by varying other individual weights in a similar manner.

If the size of the region of possible links decreases substantially, then we know the conditional independence assumption is not valid for the set of comparisons. If the conditional independence assumption were valid, then the estimated weights would accurately represent the true weights. The regions of possible links would be minimal by the theorem of Fellegi and Sunter.

A linkage rule that is based on adjusted individual comparison weights depends on the sample used in the steepest ascent procedure.

2.3.5 Specific Methods

To describe the specific methods of computing weights and obtaining corresponding linkage rules used in this paper, we need some additional background.

The only pairs considered are those that agree on at least one of the blocking criteria in Table 2.

We subdivide the set of pairs obtained via the four sets of blocking criteria into the five classes given in Table 3.

Table 2

Blocking Criteria

#	Characters Used
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.

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

Table 3

Sets of Pairs Determined by Blocking Criteria

Class	# pairs	Determining Blocking Criteria
1	1021	Agreeing on criterion 1 and no other or simultaneously agreeing on criteria 1 and 4 and no others.
2	624	Agreeing on criterion 2 and no other or simultaneously agreeing on criteria 2 and 3 and no others.
3	256	Agreeing on criterion 3 only.
4	344	Agreeing on criterion 4 only.
5	2240	Agreeing on at least one criterion but not in classes 1-4.

Class 5 contains pairs that generally agree on two or more blocking criteria. Classes 1-5 contain 2991 matches and 1494 nonmatches and miss 59 known matches. The determination of sets of blocking criteria and classes is treated in detail in Winkler (1985b, 1987).

We classify linkage rules by the different ways in which the individual comparison weights are computed and how resultant linkage rules are defined.

The first type, AA, of weight computation is an overall aggregate in all pairs. The second, A, is an overall aggregate in classes 1-4. The third, U, yields separate weight computations in classes 1-4. The fourth, C, uses steepest ascent to adjust the individual weight computation of Type U.

Each successive type of linkage rule involves increasingly more complex weight computations. Matches outside classes 1-5 are not considered in the results section because their number is constant for each of the four linkage rules.

2.4 Evaluation Procedures

The basic evaluation technique involves comparing sizes of the region of possible links when the different types of linkage rules are applied under fixed error bounds.

Efron's bootstrap (1987, 1982, 1979) is used to estimate confidence intervals for statistics such as the number of possible links. As these statistics are obtained under complicated rules, it seems unlikely that closed-form estimates can be determined.

If there are sets of pairs for which the truth and falsehood of matches are known, then we can use Efron's bootstrap to estimate the variation of parameters in the following fashion:

1. Draw calibration samples of equal size with replacement.
2. Estimate individual comparison weights of the form (2.4) using the known truth and falsehood in the sample and use them to estimate total weight via (2.3).
3. Compute cutoffs LOWER and UPPER using each sample (in our application we allow at most 2 percent of the links to be nonmatches and 3 percent of the nonlinks to be matches).
4. Using individual comparison weights from step 2, compute a total comparison weight for each pair in the entire selected set of pairs. Use cutoffs from step 2 to classify pairs as links, possible links, and nonlinks.
5. Using estimates from individual samples, determine the means and variances of the cutoff weights, of the misclassification rates, and of the number of possible links.

The bounds (2 and 3 percent, step 3) are used to try to assure that the corresponding classification error rates in the entire data base are less than 5 percent.

Table 4

Linkage Rules by Type of Weight Computation and
Sets of Pairs to Which Applied

Type	Individual Weight Computation	Linkage Rule
AA	Uniformly over all pairs in Classes 1-5	Over all pairs
A	Uniformly over all pairs in Classes 1-4	Designate pairs in Class 5 Links, Apply Fellegi-Sunter Rule to remaining pairs in Classes 1-4
U	Uniformly in each Class 1-4	Designate pairs in Class 5 Links, Apply Fellegi-Sunter Rule individually in Classes 1-4
C	Uniformly in each Class 1-4	Same as U except modify weights using steepest ascent procedure

Computations and adjustments must be performed consistently across calibration samples. Identical adjustment procedures must be used in obtaining individual adjusted weights, total weights, and cutoffs. If an individual weight is adjusted upward (step 2) by amount x or percentage y with one sample, then the same adjustment must be used with other samples.

As the underlying distributions may not be normal or may be biased and skewed, we can use new techniques of Efron (1982, 1987; also Hall 1988) to determine confidence intervals. Hall (1988) has shown the theoretical validity of the nonparametric bootstrap that includes an acceleration-constant type adjustment for skewness of a distribution.

3. RESULTS

The results in this section comprise three parts. The first part is an overall comparison from using the four different weighting methods described in section 2.3.5. The second part contains more details about the best two methods from the first part. The third part contains results from the bootstrap evaluation.

3.1 Overall Comparison

We place fixed upper bounds of 5 percent on the number of matches misclassified as nonmatches and 2 percent on the number of nonmatches misclassified as matches. As we are using discrete data, actual error rates will generally not equal their upper bounds (Table 5, columns 2 and 3).

We see that, as the complexity of the application of the weighting methodology increases, the number of possible links (size of manual review region) decreases dramatically from 1512 to 97. This indicates that the increasing complexity of the weight computations yields increasingly better decision rules.

We see that the last two methods, which both involve computing individual comparison weights separately in classes 1–4, yield the smallest sets of possible links (695 and 97, respectively).

3.2 Best Methods

We consider the best two methods, linkage rules using weights of Type U and of Type C, in greater detail. Results from applying weights of Type U and Type C are presented in Tables 6 and 7, respectively. In determining cutoff weights by class, we place rough upper bounds of 5 percent misclassified nonmatches and 2 percent misclassified matches in each class. The overall upper bound is maintained.

Comparing columns 4 and 5 across tables 6 and 7, we that the corresponding numbers of misclassified matches and nonmatches are approximately the same. This is consistent with the bounding method. In every class, the linkage rule using Type C weights yields less possible links than the rule using Type U weights.

The numbers of records classified as possible links are less in classes 1 and 4 (83 versus 55 and 44 versus 0, respectively) and dramatically less in classes 2 and 3 (409 versus 0 and 159 versus 42, respectively).

One hundred percent of the pairs in classes 2 and 4 are classified by the procedure that uses Type C weights.

Two variations distinguish the linkage rule based on type C weights from the rule based on type U weights. First, we vary agreement weights associated with the four subfields of the NAME after words have been sorted by decreasing length (Table 8). The only substantial variations (greater than 2.5 on the $\log_2$ scale) occur in Class 2.

Table 5

Error Rates and Number of Possible Links
from Applying Different Weighting Methods

Weight Type	Proportion Misclassed as		Total Classed		Possible Links
	Non-Match	Match	Non-Match	Match	
AA	.047	.020	964	2009	1512
A	.041	.015	952	2481	1052
U	.050	.020	1083	2707	695
C	.033	.019	1441	2947	97

Table 6

Results from Using a Linkage Rule Based on Type U
Weights for Delineating Matches and Nonmatches
(5 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassed as		Total Classed as		Total Not Classed	Total Records
	LOWER	UPPER	Non-Match	Match	Non-Match	Match		
1	0.5	6.5	39	14	674	264	83	1021
2	-4.5	3.5	2	4	100	115	409	624
3	-4.5	6.5	2	1	55	42	159	256
4	2.5	11.5	11	2	254	46	44	344
Totals			54	21	1083	467	695	2245

Table 7

Results from Using a Linkage Rule Based on Type C
Weights for Delineating Matches and Nonmatches
(3 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassed as		Total Classed as		Total Not Classed	Total Records
	LOWER	UPPER	Non-Match	Match	Non-Match	Match		
1	4.5	7.5	28	8	692	274	55	1021
2	2.5	2.5	5	3	379	245	0	624
3	-0.5	4.5	5	6	104	110	42	256
4	8.5	8.5	9	4	266	78	0	344
Totals			47	21	1441	707	97	2245

Table 8

Steepest Ascent Adjustment to Agreement Weights for Subfields Obtained by Wordlength Sort[1]

Class	Subfield			
	1	2	3	4
1	·	·	−	+
2	+ +	+ +	+	+
3	+	+	−	+ +
4	·	+	−	+

[1] ‘.’ means deviation less than 1.0, ‘+’, ‘−’ mean deviation greater than 1.0 and less than 2.5, and ’+ +’ means deviation greater than 2.5.

The second is that the agreement weight is only utilized if four corresponding subfields, the three subfields of CITY and the one STATE, agree. The variation, in effect, typically increases the **relative** distinguishing power of agreements/disagreements in subfields other than the CITY field.

The largest reduction (from 409 to 0) in the number of possible links takes place in Class 2. A slightly higher proportion ($.95 \approx 359/379$) of nonlinks have an agreeing CITY field than links ($.91 \approx 223/245$).

The following is an example of a match that is not designated as a link using the rule based on Type U weights but is using the rule based on Type C weights.

NAME	STREET	CITY	STATE	ZIP
Roberts Heat Oils	167 Sycamore St	Dayton	OH	53315
Maxwell S Robert Heat Oil	167 Sycamore St	Dayton	OH	53315.

The first six digits of the telephone number also agreed.

The following is an example of an erroneous match using Type C weights.

NAME	STREET	CITY	STATE	ZIP
Molar Petro	167 Sycamore St	Dayton	OH	53315
Petrochem	167 Sycamore St	Dayton	OH	53315.

These two companies do business from the same location and also have identical phone numbers.

The following is an example of an erroneous nonmatch using Type C weights.

NAME	STREET	CITY	STATE	ZIP
Johns Geo M	167 Sycamore St	Springfield	OH	53315
Geo M Johns Jobber	167 Sycamore	Spring Field	OH	53315.

Insertion or deletion of blanks in corresponding fields typically causes record pairs to be designated as a nonmatch.

Table 9

Bootstrap 90 Percent Confidence Intervals for Counts of Possible Links
500 Replications

Weight Type	Class	Ordinary Interval	BC Interval	BC_a Interval
C	1	(42,117)	(37,108)	(37,108)
C	2	(0, 0)	(7, 7)	(7, 7)
C	3	(31,154)	(34,156)	(34,156)
C	4	(0, 36)	(0, 39)	(0, 39)
U	1	(122,192)	(128,196)	(128,196)
U	2	(383,501)	(383,501)	(383,501)
U	3	(149,201)	(142,197)	(142,197)
U	4	(35, 82)	(33, 81)	(33, 81)

3.3 Bootstrap Variation

The results of this section involve increasingly more sophisticated methods of computing bootstrap confidence intervals (Table 9). For each class, 500 replications are used in computing 90 percent confidence intervals for estimates of the number of records designated as possible links. The two error bounds are fixed at 5 percent.

The first interval is the ordinary bootstrap interval that is partially based on normal theory (Efron 1979). The second interval, denoted by BC, is an interval in which a bias adjustment has been made (Efron 1979, 1982). The third interval, denoted by BC_a, is obtained using acceleration-constant type adjustments for bias and skewness (Efron 1987; also Hall 1988).

Examination of Table 9 yields that each of the intervals in respective classes are approximately the same length. If the method of adjusting to achieve weights of Type C were highly sensitive to the individual samples taken for calibration, we would expect the confidence intervals associated with Type C weights to be larger than those associated with Type U weights.

The fact that the intervals are large for either type of weight indicates the results are quite dependent on the calibrating samples. The fact that the ordinary confidence intervals are roughly the same as the BC and BC_a indicates that the respective distributions are neither biased nor skewed.

The number of possible links in intervals based on Type C weights is almost always less than the corresponding intervals based on Type U weights. Only the intervals associated with classes 3 and 4 show slight overlap. Thus, it is reasonable to accept the hypothesis that the linkage rule based on Type C weights consistently outperforms the linkage rule based on Type U weights.

4. DISCUSSION

This section is composed of four parts. The first contains a discussion of the robustness of the steepest ascent adjustments. The second subsection describes the implicit type of conditioning imposed by the steepest ascent adjustments. The third part considers the usefulness of making comparisons that are partially dependent on other comparisons. The fourth subsection describes methods for determining sets of blocking criteria.

4.1 Robustness of Steepest Ascent Adjustment

The sizes of regions of possible links are somewhat sensitive to the set of weights that are varied during the steepest ascent procedure. In two cases (one of which was presented in this paper), the numbers of possible links were approximately 100; in two others, 200. All four of the steepest ascent variations yielded improvements over the 700 possible links obtained by the best non-steepest ascent procedure.

The individual weights that were modified varied significantly over the four cases. In no case were more than eight of the 30 weights varied.

It is reasonable to hypothesize that the steepest ascent weighting procedure will yield improvements when deviations from conditional independence are substantial. No bootstrap-based significance tests were used to check the hypothesis for three of the four cases.

Obtaining small samples that allow adjustments such as performed in this paper should be straightforward. Sample sizes of 100 in each class may be sufficient. The sample sizes used for the bootstrap results of section 3.3 were approximately 100 in each class. Comparable bootstrap results using samples of 30 and 50 in each class were not sufficient to show that adjustments yielded quantifiable improvements. Sample sizes of 200 yielded bootstrap confidence intervals that were almost the same as those based on samples of sizes 100.

Many record linkage systems (*e.g.*, U.S. Dept. Agriculture 1979; U.S. Dept. of Commerce 1978a; Statistics Canada 1984) allow modification of matching parameters based on information from samples. Reestimation of parameters using sample information is a powerful feature of the Generalized Iterative Record Linkage System of Statistics Canada (1983). The parameter-reestimation in these systems generally involves direct reestimation of the marginal probabilities $m_i(\gamma^i)$ and $u_i(\gamma^i)$. It does not involve adjustments of weights such as given in this paper.

4.2 Type of Conditioning Represented by Modified Weights

To prepare for the discussion in this section, we need two sets of facts. The first set involves the conditional discriminating power of components of γ. Let σ be a vector with components $\sigma^1, \sigma^2, \ldots, \sigma^K$ that consists of a reordering of the components $\gamma^1, \gamma^2, \ldots, \gamma^K$ of γ. Then

$$\cdot\ P(\gamma|M) = P(\sigma|M) =$$

$$P(\sigma^1 = \sigma_0^1, \sigma^2 = \sigma_0^2, \ldots, \sigma^K = \sigma_0^K|M) = \tag{4.1}$$

$$P(\sigma^1 = \sigma_0^1|M) \cdot P(\sigma^2 = \sigma_0^2|\sigma^1,M) \ldots P(\sigma^K = \sigma_0^K|\sigma^1,\sigma^2, \ldots, \sigma^{K-1}, M).$$

The component σ^1 might refer to first name, σ^2 to house number, σ^3 to age, and so on.

For each σ we can call $P(\sigma^i = \sigma^{i}_{0}|\sigma^1, \sigma^2, \ldots, \sigma^{i-1}, M)$ the **successive conditional incremental discriminating component** of σ^i in M, $i = 1, 2, \ldots, K$. These incremental probability components are dependent on the reordering $\sigma^1, \sigma^2, \ldots, \sigma^K$. Each component on the right hand side of (4.1) is independent of the others. In a similar manner, we can consider incremental components in U.

The basic purpose of a reordering is to consider one specific pattern of conditional probabilities for $\gamma\epsilon\Gamma$. For the single reordering we let $\sigma = \sigma(\gamma)$ vary in $\sigma(\Gamma)$ as $\gamma\epsilon\Gamma$. Then for all $\sigma\epsilon\sigma(\Gamma)$,

$$\begin{aligned} W \equiv W(\gamma) &= \mathrm{Log}_2[m(\gamma)/u(\gamma)] \\ &= A^1 + A^2 + \ldots + A^k, \end{aligned} \tag{4.2}$$

where $A^i \equiv \mathrm{Log}_2[P(\sigma^i = \sigma_0^i | \sigma^1, \sigma^2, \ldots, \sigma^{i-1}, M)/P(\sigma^i = \sigma_0^i | \sigma^1, \sigma^2, \ldots, \sigma^{i-1}, U)]$ for $i = 1, 2, \ldots, K$.

The second set of facts involves transformations that map the ratio R given by (2.1) to real numbers which we call **weights**. For each pair of Type I and Type II errors, we consider any transformation that places weights associated with links in the highest interval, weights associated with nonlinks in the lowest interval, and weights associated with possible links in the interval between the upper and lower intervals. Such a transformation yields rules that can be represented in forms similar to form (2.2) and are equivalent to the Fellegi-Sunter rule at the same fixed pair of error levels. If the transformation is monotone, then the new weights yield rules that are equivalent to the original Fellegi-Sunter rule for all error levels.

The steepest ascent weight adjustment procedure implicitly determines a transformation of the ratio R and a single reordering that is fixed for all $\gamma \epsilon \Gamma$ and the same in M and U. The fact that the steepest ascent procedure adjusts weights sequentially assures that there is a single reordering. The adjusted weights $W^i \pm c_i$ are estimates that replace the W^i in (2.3) for some real constants c_i, $i = 1, 2, \ldots, k$.

The fact that the adjusted weights yield smaller regions of possible links means that, at a fixed pair of error levels, the new total weights more accurately represent a transformation of the Log_2 of the ratio of the true probabilities given by the left hand of (4.1). The new total weights represent estimates that transform the right hand side of (4.2).

The adjustment procedure allows us to utilize better the incremental distinguishing power of one field given another, a second field given the first two, and so on. We note that we do not need to know the specific transformation or the specific pattern of conditioning induced by the reordering.

The adjustment procedure is similar to new bootstrap procedures (Efron 1987; Hall 1988). The validity of the bootstrap procedures is dependent on the existence of monotone transformations, bias constants, and acceleration constants that yield the exact correspondence of confidence intervals of the original distributions with confidence intervals of specified normal distributions. The transformations and constants need not be known.

4.3 Value of Dependent Comparisons

The intuitive idea of making a number of comparisons, some of which may be partially dependent on other comparisons, is that they may, when used in properly adjusted rules, yield additional distinguishing power. Newcombe and Kennedy (1962, see also Newcombe *et al.* 1983) have given examples of comparisons of portions of name fields that intuitively may be dependent on other comparisons. The additional comparisons, nevertheless, may yield better linkage rules than those rules that do not utilize the same additional comparisons.

The chief difficulty in using additional comparisons is properly utilizing their incremental distinguishing power. This paper's set of comparisons - in particular, of subfields of the name field - is not independent in the sense of equation (2.3). The primary purpose of the set is to illustrate methods for systematically obtaining better linkage rules when the conditional independence assumption is not valid.

4.4 Additional Blocking Criteria

There are two conflicting goals when a set of blocking criteria is used to reduce the number of pairs in $A \times B$ that receive further processing. The first is the need to reduce (drastically) the number of pairs that are processed and to obtain a set in which linkage rules can accurately delineate matches and nonmatches. The second is to obtain a set that contains as many matches from M as possible.

To determine whether it is feasible to look for additional sets of blocking criteria, it is first necessary to find estimates of the number of matches missed by a given set of blocking criteria. If the estimates are acceptably small, then it is not necessary to look for additional criteria.

To estimate the number of matches missed by given sets of blocking criteria, Scheuren (1983) suggested using standard capture-recapture techniques such as given in Bishop, Fienberg, and Holland (1975, Chapter 6). Winkler (1987) applied the techniques to the same empirical data and four sets of blocking criteria as in this paper.

The best fitting loglinear model for the table of counts of records captured and not captured by the four sets of blocking criteria was used in obtaining a confidence interval for the number of matches missed. Based on assumed asymptotic normality, a 95 percent confidence interval (27,160) was computed. The interval represents between 1 and 5 percent of the matches.

5. SUMMARY

The results of this paper show that the conditional independence assumption is not always valid. When the assumption is not valid, it is possible to develop adjusted linkage rules that improve on the standard linkage rule. Under fixed bounds on error rates, the improved rules reduce the size of the region of possible links.

REFERENCES

BISHOP, Y. M. M., FIENBERG, S. E., and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis.* Cambridge: MIT Press.

COCHRAN, W.G., and COX, G.M. (1957). *Experimental Designs.* New York: John Wiley and Sons.

EFRON, B. (1979). Bootstrap methods: Another look at the Jackknife. *Annals of Statistics,* 7, 1-26.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Methods.* Philadelphia: SIAM.

EFRON, B. (1987). Better Bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association,* 82, 171-185.

FELLEGI, I. P., and SUNTER, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association,* 40, 1183-1210.

HALL, P. (1988). Theoretical comparison of Bootstrap confidence intervals. *Annals of Statistics,* 16, 927-953.

KELLEY, R. P. (1986). Robustness of the Census Bureau's record linkage system. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 620-624.

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 Association for Computing Machinery,* 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. *Computers in Biology and Medicine,* 13, 157-169.

SCHEUREN, F. (1983). Design and estimation for large federal surveys using administrative records. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 377-381.

SCHEUREN, F. (1985). Methodological issues in linkage of multiple data bases, in *Record Linkage Techniques - 1985,* edited by W. Alvey and B. Kilss, U. S. Internal Revenue Service, Publication 1299 (2-86), 155-167.

STATISTICS CANADA (1982). Record Linkage Software, Systems Development Division.

STATISTICS CANADA (1983). Generalized Iterative Record Linkage System, Systems Development Division.

STATISTICS CANADA (1984). Record Linkage Software, EDP Planning and Suport Division.

TEPPING, B. J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association* 63, 1321-1332.

U.S. DEPARTMENT OF AGRICULTURE (1979). List Frame Development: Procedures and Software, Statistical Reporting Service.

U.S. BUREAU OF THE CENSUS (1978a). UNIMATCH: A Record Linkage System, Survey Research Division.

U.S. BUREAU OF THE CENSUS (1978b). ZIPSTAN: Generalized Address Standardizer, Survey Research Division.

WINKLER, W. E. (1985a). Preprocessing of lists and string comparison, in *Record Linkage Techniques - 1985*, edited by W. Alvey and B. Kilss, U. S. Internal Revenue Service, Publication 1299 (2-86), 181-187.

WINKLER, W. E. (1985b). Exact matching lists of businesses: Blocking, subfield identification, and Information Theory, in *Record Linkage Techniques - 1985*, edited by W. Alvey and B. Kilss, U. S. Internal Revenue Service, Publication 1299 (2-86), 227-241.

WINKLER, W. E. (1985c). Exact matching lists of businesses. *Proceedings of the Section on Survey Research Methods, American Statistical Association* 438-443.

WINKLER, W. E. (1987). An Application of the Fellegi-Sunter Model of Record Linkage to Lists of Businesses, Energy Information Administration, Technical Report.

WINKLER, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter Model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association* to appear.