Survey Methodology, June 1993
Vol. 19, No. 1, pp. 13–29
Statistics Canada

Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment

THOMAS R. BELIN[1]

ABSTRACT

Record linkage refers to the use of an algorithmic technique for identifying pairs of records in separate data files that correspond to the same individual. This paper discusses a framework for evaluating sources of variation in record linkage based on viewing the procedure as a "black box" that takes input data and produces output (a set of declared matched pairs) that has certain properties. We illustrate the idea with a factorial experiment using census/post-enumeration survey data to assess the influence of a variety of factors thought to affect the accuracy of the procedure. The evaluation of record linkage becomes a standard statistical problem using this experimental framework. The investigation provides answers to several research questions, and it is argued that taking an experimental approach similar to that offered here is essential if progress is to be made in understanding the factors that contribute to the error properties of record-linkage procedures.

KEY WORDS: Cutoff weight; False-match rate; Fellegi-Sunter algorithm; Matching variables; Post-enumeration survey; String comparison; Weighting scheme.

1. EVALUATING RECORD-LINKAGE PROCEDURES

Record linkage refers to the use of an algorithmic technique to identify pairs of records, one from each of two data files, that correspond to the same individual. The goal is to identify, using a computerized approach, the records from the respective data files that should be declared "matched" as well as the records that should be declared "not matched" without an excessive rate of error, thereby avoiding the cost of manual processing.

Specifying a record-linkage procedure requires both a method for measuring closeness of agreement between records and a rule for deciding when to classify records as matches or non-matches. Much attention has been paid in the record-linkage literature to the problem of assigning so-called "weights" to individual fields of information in a multivariate record to obtain a "composite weight" that summarizes the closeness of agreement between two individuals (*e.g.*, Newcombe *et al.* 1959; Fellegi and Sunter 1969; Newcombe 1988; Copas and Hilton 1990). Less attention has been paid to other aspects of record-linkage procedures, such as the handling of close but inexact agreement between fields of information, and to the effects of using various approaches (treatments) in combination with one another.

In some settings, a personal identifier, such as a social security number, can serve as a basis for linkage. However, such an identifier is not always available, and even when one is present, it still may be necessary to rely on other identifying information for a substantial subset of cases (*e.g.*, Rogot, Sorlie and Johnson 1986).

This paper describes a large factorial experiment contrasting various procedures for matching census and post-enumeration survey (PES) records. Social security number is not collected in the census, so we are in a setting where closeness of agreement is based on several variables. Interest focuses on two questions:

(1) What are the most important factors affecting the accuracy of record linkage?

(2) What combination of factors works best in practice?

Beyond addressing these questions in the census/PES setting, perhaps the most important contribution of this investigation is the idea that record-linkage procedures should be studied by conducting careful experiments. With many factors at the discretion of the operator of the program, there is little hope of understanding the full complexities of a matching algorithm by varying factors one at a time (or worse, not even conducting any systematic evaluation at all). The idea of conducting an experiment would seem quite natural to an agricultural scientist or an industrial quality-control engineer, although it seems that such an approach has not been taken in the context of record linkage aside from this investigation and earlier work by the author (Belin 1989a, 1989b).

2. APPLIED CONTEXT FOR RECORD LINKAGE

2.1 Applications of Record Linkage

Record-linkage methods have been used in a variety of settings. Applications can be characterized as falling into two broad groups: problems where it is desired to draw

[1] Thomas R. Belin, Department of Biomathematics, UCLA School of Medicine, Los Angeles, CA, 90024-1766, U.S.A.

inferences about relationships between variables collected in separate large data files, and problems where interest focuses directly on the number of individuals represented in one or both data files (or a function of those quantities).

Examples of the first type of application are numerous. Studies have been conducted linking data from health and nutrition surveys to registries of mortality data to study relationships between dietary risk factors and death from various causes (Johansen 1986), linking labor force survey data to mortality data to assess health effects of uranium mining (Newcombe, Smith, Howe, Mingay, Strugnell and Abbatt 1983; Abbatt 1986), linking information on educational background to records of earnings of individuals some years later to assess the benefit of a college education (Fagerlind 1975), comparing reported income on welfare records to reported income on tax records (Kershaw and Fair 1979), and linking records of individuals exposed to radiation during atomic-bomb tests and records of a cohort of control individuals to national death records to assess differences in mortality patterns between exposed and control individuals (Dulberg, Spasoff and Raman 1986). Using record-linkage methodologies in such studies is attractive primarily for reasons of cost and timeliness, since for any of the research endeavors just described, it would take much longer and would have been much more expensive to conduct studies with one or more stages of followup than it was to make use of existing data.

The primary motivating example in this article is representative of the other type of application, where the goal is to determine the number of overlapping cases in two data files. In this example, a record-linkage procedure is used as the first step of an extensive matching operation in which records from a census are compared to records from a large-scale post-enumeration survey (PES) conducted after the census to evaluate census coverage. Other examples where the goal is to determine the number of overlapping cases between data files are the investigation by Nicholl (1986) of classification errors regarding the types of injuries sustained by road accident victims (based on linking hospital records to police reports of accidents), the investigation by Johnson (1991) into caseloads for U.S. Attorneys in different districts around the country (based on linking a list of cases assembled by the Department of Justice to a list of cases assembled by federal district courts), and a variety of investigations into the accuracy and coverage of mortality data files (Wentworth *et al.* 1983; Curb *et al.* 1985; Boyle and Decouflé 1990; Williams *et al.* 1992).

Census undercount estimation has been a prominent and at times controversial topic in statistical research, especially during the past decade. Much of the controversy revolves around a proposed adjustment of the census based on undercount estimates from a PES. For general background on issues involved in census undercount estimation, see Ericksen and Kadane (1985), Citro and Cohen (1985), Freedman and Navidi (1986), Wolter (1986), Schirm and Preston (1987), Ericksen, Kadane, and Tukey (1989), Cohen (1990), and the special sections on census coverage error in the June and December, 1988, issues of this journal. A record-linkage procedure is the first step of matching census records to PES records; it is followed by matching of records by clerks, subsequent followup interviewing of households when there appear to be discrepancies between the census and PES findings, and an additional round of clerical matching after followup interviewing. Based on assessments from the matching operation and certain assumptions about the probability that individuals would be included only in the census, only in the PES, in both the census and PES, or in neither the census nor PES, it is possible to estimate undercount (or overcount) rates in the census.

2.2 Background on Record-Linkage Theory

The development probabilistic reasoning in record-linkage theory can be traced to Newcombe, Kennedy, Axford, and James (1959), who develop a weighting scheme in an effort to reflect the odds that a pair of records is correctly matched. Fellegi and Sunter (1969) enhance the theoretical underpinnings of commonly-used weighting rules, noting that the procedure proposed by Newcombe *et al.*, corresponds to calculating a likelihood ratio under a simple model for the record-linkage problem that supposes independence of agreement among all fields of information within records. They show that a weighting scheme similar to that of Newcombe *et al.*, combined with cutoff weights that depend on a specified false-match rate and a specified false non-match rate, define a linkage procedure that is optimal in the sense of minimizing the proportion of records that will be assigned neither as definitely matched nor as definitely not matched, assuming the underlying model is valid.

Much of the ensuing development of record-linkage technology has taken place in the context of applications, as investigators put the theoretical ideas outlined in the earlier literature to practical use. Prominent applications include the Oxford Record Linkage Study (Acheson 1967; Goldacre 1986); the three-way match among records from the Current Population Survey, the Social Security Administration, and the Internal Revenue Service (Kilss and Scheuren 1978); and the National Longitudinal Mortality Study (Rogot, Sorlie, Johnson, Glover and Treasure 1988). The proceedings volumes from conferences on record linkage (Kilss and Alvey 1985; Howe and Spasoff 1986; Carpenter and Fair 1990), compilations of papers from annual conferences (Kilss and Alvey 1984a; Kilss and Alvey 1984b; Kilss and Alvey 1984c; Kilss and Alvey 1987; Kilss and Jamerson 1990), and proceedings volumes from conferences more broadly focused on uses of administrative

data (Coombs and Singh 1988) document numerous other applications that make use of record-linkage methodology.

Software development has enhanced the ability to pursue research into record linkage. Software incorporating refinements of weighting methods and blocking strategies has been developed for use in a variety of applications at Statistics Canada and the U.S. Bureau of the Census. Background on the Statistics Canada "Generalized Iterative Record Linkage System" (GIRLS) is discussed in Howe and Lindsay (1981); documentation is contained in Hill (1981) and Hill and Pring-Mill (1986). Background on the matching system developed by the Record Linkage Staff at the U.S. Bureau of the Census can be found in Jaro (1989), Winkler (1989), and Winkler and Thibaudeau (1992), with documentation found in Laplant (1988), Laplant (1989), and Winkler (1991).

New models that reflect subtleties within data files that could be used in developing a probabilistic weighting scheme are offered by Copas and Hilton (1990). Other extensions to record-linkage methodology designed to take advantage of information in person names are described in Newcombe, Fair and Lalonde (1992). A review paper by Jabine and Scheuren (1986), a textbook by Newcombe (1988), and a compilation by Baldwin, Acheson and Graham (1987) serve as broad references on record-linkage methodology.

2.3 Flow of a Standard Record-Linkage Procedure

Typical steps in a record linkage procedure can be described as follows: (1) data collection, (2) preprocessing of data, (3) determination of rules for assessing closeness of agreement between candidate matched pairs, (4) assignment of candidate matched pairs, and (5) declaration of matched pairs. We use the term "candidate matched pairs" to describe pairs of records that are brought together as being the best potential match for each other from the respective data files (*cf.* "hits" in Rogot, Sorlie, and Johnson (1986); "pairs" in Winkler (1989); "assigned pairs" in Jaro (1989)). Candidate matched pairs might be declared matched after the application of a decision rule in step (5), but they will not necessarily be declared matched by the decision rule.

As indicated earlier, closeness of agreement between candidate matched pairs is assessed in many record-linkage procedures by a univariate summary statistic, often referred to as a "composite weight". In such procedures, step (3) above would refer to the determination of weighting rules, and step (5) above would involve the setting of a cutoff weight above which record pairs will be declared matched.

Record linkage may be viewed as a decision problem with two or more actions to be taken by the computer. Typically, three actions are considered (*e.g.*, declare records matched, declare records as not matched, or send record to be reviewed more closely by a human observer, as in Fellegi and Sunter 1969), although sometimes only two actions (declare matched, declare not matched) are contemplated, and as many as five actions have been considered in some instances (Tepping 1968).

Postulating that distance between multivariate records can be summarized by a univariate composite weight narrows the scope of possible procedures that could be used to perform record linkage. The author is aware of very little research exploring alternatives to such univariate-composite-weight approaches, other than merely specifying a deterministic set of rules for when to declare records matched; one exception is Smith and Newcombe (1975). Such alternatives are beyond the scope of this paper.

2.4 Detailed Description of the Procedure Used to Match Census/PES Records

A variety of separate techniques may be involved in each of the five steps outlined above. Figure 1 provides a flowchart illustration of the main steps used in the linkage of census/PES records.

The frame of the census is a compilation of housing-unit address listings. Addresses are assembled by a variety of techniques, generally depending on whether the area is urban or rural. In urban and suburban areas, census forms are mailed to households with the hope that residents will respond by mailing back a completed form; in other areas census enumerators visit households. When there is no response from a household that was sent a census form by mail, an enumerator will visit the household in person. Data are entered into Census Bureau computer files by a combination of computerized scanning techniques and clerical keying operations. An overview of census methodology can be found in Citro and Cohen (1985); detailed descriptions of various census operations can be found in the Census Bureau's 1990 Decennial Census Information Memorandum Series (Bureau of the Census 1988-1991).

Data collection in the type of post-enumeration survey conducted in 1990 (and in test censuses leading up to the 1990 PES) begins with a process of listing addresses that is conducted by enumerators canvassing neighborhoods. Information is obtained entirely through interviewing operations as opposed to the mailout-mailback approach. Data are entered into computer files entirely by clerical keypunching. Hogan (1992) provides an overview of the PES; details of PES operations can be found in the Census Bureau's STSD Decennial Census Memorandum Series (Bureau of the Census 1987-1991).

Preprocessing of data is rarely discussed in the literature on record linkage, even though this stage provides opportunities both for squeezing available information from the data at hand and for unwisely discarding information available from the data. Winkler (1985a, 1985b) presents

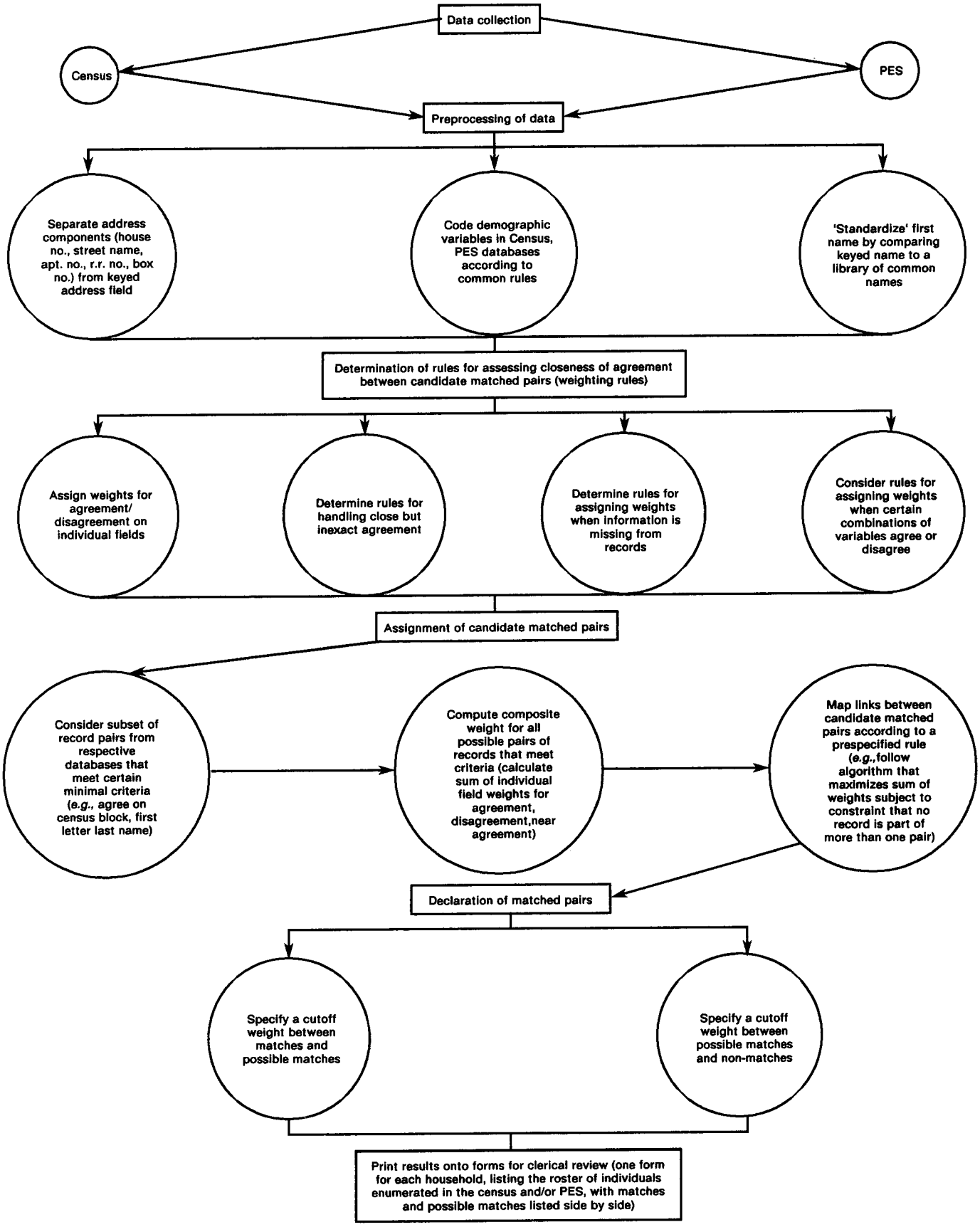

Figure 1. Flowchart of Census/PES Record Linkage Procedures

some specific strategies that are shown to make it easier to distinguish true matches from false matches, and Jabine and Scheuren (1986) and Newcombe (1988) offer some broad guidelines in this area. In the census/PES matching operation, preprocessing of data includes coding demographic variables according to common rules, identifying and separating address components (such as house number, street name, apartment number, rural route number, and post office box number) from the keyed address field (Laplant 1989), and "standardizing" an individual's first name by comparing the keyed first name to a library of nicknames and converting nicknames observed in the data to their common antecedent names (Paletz 1989).

The census/PES record-linkage procedure is a weight-based procedure. The determination of a weighting method includes consideration of both model-based and *ad hoc* rules for assigning weights for agreement and disagreement on individual fields of information, rules for assigning weights for close but inexact agreement on particular fields, rules for assigning weights when information is missing from records, and rules for assigning weights when certain combinations of variables are found to be in agreement or disagreement.

The designation of candidate matched pairs in census/PES matching reflects certain constraints that are placed on the matching process. First, time and resource constraints make it impractical to compare each record in one data file to every record in the other data file. Accordingly, comparisons are made only between pairs of records that meet certain minimal criteria, such as that they fall in the same census block and share the same first letter of last name. The subset of records formed by this restriction is referred to as a "block", and the variables required to be in agreement for a match to be declared are referred to as "blocking variables" (Jaro 1989).

Another constraint placed on the census/PES matching operation is that a given record in one data file is not allowed to be declared matched to more than one record in the other data file. The approach that is used to perform the assignment of candidate matched pairs draws on operations-research techniques for solving the so-called transportation problem (Jaro 1989). The algorithm assigns candidate matches so as to maximize the sum of composite weights among all possible pairs of records within a block defined by the blocking variables, subject to the aforementioned restriction that no record is allowed to match more than one record in the other data file. For example, suppose that within a particular block record A from file 1 has a higher agreement weight with record B from file 2 than with any other record in file 2. The assignment algorithm still might link record A to another record, say C, and link B to another record, say D, if the sum of the agreement weights for (A,C) and (D,B) are higher than for other permutations of candidate match assignment.

The current approach to census/PES matching contemplates three possible actions to be taken by the computer: declare a record pair to be a match, declare a record pair to be a "possible match", or declare a record to be not matched. All non-matches and possible matches are sent to clerks to be reviewed, and an attempt to obtain a followup interview is made for households where there is a discrepancy between the census and the PES. The distinction between possible matches and non-matches only has to do with the procedures applied by clerks when they review these cases (Childers 1989; Donoghue 1990). In the processing of 1990 census/PES data, the operator of the matching program set cutoff weights manually to distinguish matches, possible matches, and non-matches after scanning sets of candidate matched pairs with weights in a certain range. A new technique by Belin and Rubin (1991) offers an alternative for automating the setting of cutoffs.

3. AN EXPERIMENT

3.1 Factors Influencing the Output of Record-Linkage Procedures

The performance of a record-linkage procedure can depend on a number of factors, including:

(1) The choice of matching variables;

(2) The choice of blocking variables;

(3) The assignment of weights to agreement or disagreement on various matching variables;

(4) The handling of close but not exact agreement between matching variables;

(5) The handling of missing data in one or both of a pair of records;

(6) The algorithm for assigning candidate matches;

(7) The choice of a cutoff weight above which record pairs will be declared matched;

(8) The site or setting from which the data are obtained.

Among these factors, only (8) represents a source of variation over which the operator of the matching program does not have control. As mentioned earlier, two lines of inquiry are of primary interest in the experiment. Identifying major sources of variability in record linkage could help to focus future record-linkage research and to offer a deeper understanding of the process that generates errors in linkage procedures. Further, it is of interest to identify the combination of factors that works best in achieving a maximum number of matches while maintaining low error rates, since in practice the user generally must make a single choice among a myriad of possibilities for each factor just described.

3.2 Factorial Experiment Using Census/ Post-Enumeration Survey Data

A study was conducted using data from each of the three sites (St. Louis, Missouri; East Central Missouri including the Columbia, Missouri area; and a rural area in eastern Washington state) of the 1988 dress rehearsal census and PES. These data sets had been matched by computer and then reviewed by clerks. For the purposes of subsequent analysis, the final clerical determinations of true and false match status are taken as the truth. Thus, although subsequent analyses will only be as accurate as the determinations by clerks, these data files offer an excellent opportunity to study record linkage.

Descriptions of the specific methods used in linking records between the census and PES can be found in Jaro (1989), Winkler (1991), and Winkler and Thibaudeau (1992). The current implementation of the record-linkage procedure allows the user a variety of options over all of the factors listed in Section 3.1 except for the choice of an algorithm for assigning candidate matches (a "linear-sum assignment" algorithm is used; see Jaro 1989).

The variables available for matching census/PES records include name, address, age, race, sex, telephone number, marital status, and relationship to head of household. In practice, name is usually broken down into first name, last name, and middle initial, with these three used as separate matching variables. A preprocessing program is typically used to parse address information into house number, street name, apartment number, rural route number, and box number (Laplant 1989). Sometimes "irregularities" in address information, perhaps caused by clerical typing errors or by recording errors on the part of a census or post-enumeration survey interviewer, result in an inability to parse an address into various components; in these cases, the entire address field (referred to as the "conglomerated address") is used as a matching variable. An available preprocessing program also can be used to convert nicknames to a "standardized" name using a library of names and their common variants (Paletz 1989). A variety of schemes are available for assigning weights based on close agreement between variables, and a procedure is also available for adding or subtracting weight to the composite weight for a record pair when certain combinations of fields are in agreement or disagreement (Winkler 1991).

The experiment consisted of eight "treatment" factors and one "blocking" factor (where "blocking" here refers to the experimental-design notion of a grouping of units expected to yield results as similar as possible in the absence of treatment effects) with replication across three sites in a $2^5 \times 3^3 \times 5 \times 13$ factorial design. The outcome variable in the experiment, described further in Section 3.5, was a transformation of the false-match rate, where the transformation was used to stabilize the variance of the outcome. The factors in the experiment can be described as follows:

Label	Description of factor	Number of levels of factors	Description of levels of factor
A	Assignment of weight for name fields.	5	1. Assign weights of ± 2 for agreement/ disagreement on first, last name. 2. Assign weights of ± 4 for agreement/ disagreement on first, last name. 3. Assign weights of ± 6 for agreement/ disagreement on first, last name. 4. Assign weights based on estimates of probabilities of agreement on first, last name from Fellegi-Sunter algorithm (see Winkler and Thibaudeau 1992). 5. Use frequency-based weighting for first, last name (see Winkler and Thibaudeau 1992).

Label	Description of factor	Number of levels of factors	Description of levels of factor
B	Assignment of weight for close but inexact agreement on name fields.	3	1. Assign disagreement weight for any discrepancy in first, last name. 2. Assign fraction of agreement weight for close agreement on first, last name using Jaro string comparison metric (Jaro 1989; Winkler 1991). 3. Assign fraction of agreement weight for close agreement on first, last name using piecewise linear metric described in Winkler (1991).
C	Assignment of weight for non-name fields.	2	1. Assign weights of ± 2 for agreement/ disagreement on age, phone number, and address fields, and assign weights of ± 1 for agreement/ disagreement on sex, race, marital status, relationship to head of household, middle initial. 2. Assign weights based on estimates of probabilities of agreement from Fellegi-Sunter algorithm.
D	Assignment of weight for close but inexact agreement on non-name fields.	3	1. Assign disagreement weight for any discrepency in non-name fields. 2. Assign fraction of agreement weight for close agreement on house number, street name, phone number, age, using Jaro string comparator. 3. Assign fraction of agreement weight for close agreement on street name using Jaro string comparator, for age using Jaro pro-rated-to-absolute-difference metric, for house number and phone number using Winkler piecewise-linear string comparator.
E	Use of keyed first name or standardized version of first name.	2	1. Use the version of the individual's first name that was keyed into each data file for comparison of first name. 2. Use the version of the individual's first name that is obtained as output from name standardization software (Paletz 1989).

Label	Description of factor	Number of levels of factors	Description of levels of factor
F	Adjustment of weights for correlated agreement.	2	1. Do not adjust the composite weight for possible correlated agreement. 2. Adjust composite weights for possible correlated agreement between first name, middle initial and among first name, sex, age.
G	Inclusion of marital status, relationship to head of household as matching variables.	2	1. Do not include marital status, relationship as matching variables. 2. Include marital status, relationship as matching variables.
H	Use of four or seven digits of phone number.	2	1. Use only last four digits of phone number as a matching variable. 2. Use all seven digits of phone number.
I	Site of census/post-enumeration survey.	3	1. Eastern Washington state. 2. Columbia, Missouri. 3. St. Louis, Missouri.
J	Proportion of PES file declared matched.	13	1.-13. Let the number of records accepted as declared matches equal 60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5%, 90% of the number of PES records in the given site.

With reference to the sources of variation described in Section 3.1, factors E, G, and H relate to the choice of matching variables; factors A, C, and F relate to the choice of a weighting scheme; factors B and D relate to the handling of close but inexact agreement; factor J reflects the choice of a cutoff; and factor I reflects the influence of the particular site on the performance of the matching procedure.

Consideration of resource limitations led to a decision not to address the effect of varying missing data treatments or the effect of different choices of blocking variables in this experiment, and the lack of available software precluded any investigation of alternative alogrithms for assigning candidate matches. Belin (1989a, 1989b) studied the influence of missing data treatments and of different choices of blocking variables in an experiment similar to the factorial experiment described here. The results of that investigation suggested that alternative treatments of missing data had no substantial effect on false-match rates associated with different cutoffs in matching of census/PES data, but the choice of blocking variables did have a substantial effect.

In this investigation, as in Belin (1989a, 1989b), only "one-pass" matching procedures are considered. That is, the entire computer-matching operation consists of a single cycle of choosing blocking variables, establishing weights, and setting a cutoff, as opposed to "multiple-pass" procedures that first use very restrictive blocking variables to skim off the nearly perfect matches, then relax the blocking criteria in successive passes through the data. The author is aware of very little research on multiple-pass matching procedures. Belin (1989b) reports that when single-pass procedures are used, procedures that use relatively less restrictive blocking criteria enjoy advantages over procedures that use relatively more restrictive blocking criteria, confirming the intuitive notion that the blocking process can exclude true matches from consideration as an unfortunate side effect.

3.3 Subtleties in Experimental Treatments

3.3.1 Treatments for Assigning Weights for Agreement/ Disagreement on Fields of Information

To clarify the experiment, we describe each of the experimental factors in greater detail. Factors A and C are concerned with the assignment of weights for agreement and disagreement on the various matching variables. The different weighting approaches used in factors A and C include completely *ad hoc* methods and methods that are based on estimates of parameters in explicit probability models. The study of *ad hoc* weights provides an opportunity to gauge the importance of incorporating more complicated approaches to weighting.

The *ad hoc* weighting schemes call for a weight of U, say, to be added to the composite weight if the fields being compared agree, and for an identical weight U to be subtracted from the composite weight if the fields being compared disagree. Three different values of U are studied in factor A, with the same value of U being assigned for agreement on first name as for agreement on last name. In factor C, an *ad hoc* scheme that weights some variables more than others is studied, with the decision about which variables to weight more being based on *a priori* judgments. Belin (1989b) suggests that such a "modified-equal-weighting" scheme has advantages over an "equal-weighting" scheme in which all matching variables are assigned the same weights for agreement or disagreement.

The "Fellegi-Sunter algorithm" refers to the method outlined in Fellegi and Sunter (1969), which is based on a probabilistic model that incorporates information about patterns of agreement and disagreement between pairs of records. The model postulates that probabilities of agreement on individual fields of information given that a pair is a true match are independent across all fields of information, and that independence across fields also holds given that a pair is a false match. The paper by Fellegi and Sunter shows that such a model implies certain optimality properties for the type of weighting scheme used by Newcombe *et al.* (1959), in which weights for individual fields of information are calculated by taking the logarithm of the ratio of probability of agreement given true match to the probability of agreement given false match, and in which composite weights are obtained by summing individual field weights.

In applications, the probabilities of agreement given true match and agreement given false match need to be estimated. For the treatments in the experiment characterized as relying on the Fellegi-Sunter weighting approach, the probabilities of agreement given true match are estimated using a version of an EM algorithm (Dempster, Laird and Rubin 1977) to obtain maximum likelihood estimates of these probabilities based on counts of all possible patterns of agreement observed in the data files at hand (Winkler 1989; Jaro 1989). The probabilities of agreement given false match are estimated based on counts of agreement on individual fields between all record pairs that agree on blocking variables, making use of the fact that most of the pairs that could possibly be brought together as matches are not true matches (Winkler and Thibaudeau 1992).

Another weighting approach that has been implemented in the Census Bureau's record linkage software considers the relative frequency of names in the data files at hand, assigning more weight for agreement on names such as Abramowicz, which may be relatively rare, than for agreement on names such as Smith, which may be common. Of course, it could happen that in a particular area Abramowicz is a more common name than Smith, in which case the frequency-based weighting approach would assign greater weight to agreement on the name Smith. The idea of incorporating information on marginal frequencies from the current data files was mentioned by Newcombe *et al.* (1959), and has been noted by many authors since then, including Fellegi and Sunter (1969). (Thus, the distinction drawn here between the "Fellegi-Sunter algorithm" and "frequency-based weighting" is actually a distinction between two methods of calculating weights that are both discussed by Fellegi and Sunter.) Details on the implementation of frequency-based weighting in the Census Bureau's software can be found in Winkler and Thibaudeau (1992).

3.3.2 Treatments for Handling Close but Inexact Agreement

Factors B and D deal with the handling of fields that may agree closely but do not agree exactly with one another. Several techniques have been proposed for handling close but inexact agreement between fields of information, often reflecting different perspectives on probable departures from exact agreement.

The Jaro string comparator is designed to measure the closeness of agreement of two multi-character fields; the metric that defines closeness is a function of the lengths of the character fields in the two files, the number of characters in common between the character fields, and the number of transpositions of characters between the character fields. The weight that gets assigned for partial agreement is between the weight for agreement on the field and the weight for disagreement on the field, and is a linear function of the string comparator metric between the agreement weight and the disagreement weight.

The Winkler piecewise-linear approach uses the same metric as the Jaro string comparator to define closeness of agreement, but the rate at which partial agreement weights decrease from the agreement weight to the disagreement weight is a piecewise linear function of the string comparator metric, requiring two user-supplied rate parameters and two user-supplied thresholds where the slope changes.

The Jaro pro-rated method assigns a weight between the agreement weight and the disagreement weight based on the absolute value of the difference between two numeric fields. As with the aforementioned techniques, the partial agreement weight falls off as a linear function of the absolute value of the difference.

Even for some numeric fields (*e.g.*, telephone number), a comparison method designed to accommodate slight typographical variation would seem more sensible than a method based on absolute numerical difference. However, for variables such as year of birth or age, it may not be clear whether to target efforts toward accommodating typographical errors (for which a string comparison method would be best suited), reporting errors (for which the absolute-difference method may be most appropriate), or other types of errors such as "heaping" or rounding of reported ages on multiples of five years (for which neither of the previously mentioned comparison methods would be ideally suited). Accordingly, we pursue our empirical evaluations in an attempt to shed light on these issues.

3.3.3 Treatments Involving the Choice of Matching Variables

As mentioned previously, an approach has been developed at the Census Bureau for converting nicknames to a standardized root. Software developed by Paletz (1989) implements the name-standardization routine.

The treatment that omits marital status and relationship to head of household as matching variables allows for an assessment of the importance of two background demographic variables on the quality of matching. Chernoff (1980) develops theory for the information carried by a matching variable and shows that a variable recorded in error even a small percentage of the time can lose a substantial amount of information for matching purposes (*e.g.*, the Kullback-Leibler information associated with a binary variable recorded in error three percent of the time is only about half that of a binary variable recorded without error). Considering that relationship to head of household could differ between the census and PES if the person listed as the head of household is different, and that marital status will change for some individuals in the intervening time, it is not clear in advance how much information for matching is provided by these variables. On the other hand, it is hard to imagine that using additional matching variables would be deleterious, so that this treatment provides a standard for assessing the practical significance of some of the other treatments.

The treatment of using either four or seven digits of phone number as a matching variable is self-explanatory. A motivation for considering this treatment is that one of the specific piecewise-linear string comparator methods proposed by Winkler was developed based on analysis of the last four digits of phone number as a matching variable.

3.3.4 Treatment for Adjusting Composite Weights for Correlated Agreement

The method described as adjusting the composite weight to reflect the possibility of correlated agreement is also due to Winkler and is described in Winkler and Thibaudeau (1992). Research by Kelley (1986) and Thibaudeau (1989) reveals that agreement on the various fields available for matching between the census and PES data files is far from being independent across fields. In particular, analyses suggested that agreement on first name was correlated with agreement on middle initial and that agreement on first name, age, and sex were mutually correlated. These findings led to the implementation of modifications to the composite weight when certain patterns appear (*e.g.*, if first name, age, and sex all disagree, then a large value is subtracted from the composite weight). The current scheme for adjusting the composite weight is entirely *ad hoc*; research into methods that reflect correlated agreement still appears to be in its infancy.

3.4 Data Files Used in Experiment

As mentioned before, the three sites of the 1988 dress rehearsal census and post-enumeration survey provided separate data files on which these analyses of record linkage could be performed. There were 12,072 records in the PES file from St. Louis, 6,581 records in the PES file from East Central Missouri, and 2,782 records in the PES file from eastern Washington state. As was also noted earlier, the final determinations by clerks who reviewed these files were taken as the truth for purposes of evaluation. Other test censuses were conducted during the 1980's; the primary reason for not including the data from other test censuses in this experiment is that a considerable amount of "overhead" time is required to prepare a data set for the analyses performed here.

3.5 Outcome Variable

The primary outcome variable considered in this experiment was a transformation of the false-match rate. The false-match rate is defined as the number of false matches divided by number of declared matches, and is a common measure of performance in the literature on record linkage (*e.g.*, Fellegi and Sunter (1969) attempt to provide output that satisfies a fixed false-match rate criterion supplied by the operator of the program). In order to stabilize the variance of the outcome, the analyses here use the arcsine of the square root of the false-match rate as an outcome variable.

3.6 Choice of Cutoff Weight as a Blocking Factor

It is clear that the false-match rate in record linkage is apt to depend heavily on the choice of a cutoff between declared matches and declared non-matches. Accordingly,

a blocking factor (Factor J) is introduced to fix the determination of cutoffs so as to facilitate comparison of other record-linkage treatments. To provide a standard for comparisons across sites having different numbers of records, the cutoff level is defined in terms of the proportion of the PES data file declared matched.

Because of the discreteness of record-linkage weights, it is possible to have ties among the weights of record pairs on the boundary where the cutoff should be assigned. For example, in a file of 10,000 records, there may be 40 records with weight W (of which 10 may be false matches), 7,980 records with weight greater than W (of which 3 may be false matches), and 1,980 records with weight less than W. If the treatment in factor J calls for 80% of the PES file to be matched, then it may not be obvious how to calculate the false-match rate, since there are 40 records with the same weight straddling the point where the cutoff should be set. Calculations of the false-match rate in such a case are based on the following relationship:

$$\text{fmr} = \frac{f_{\text{abv}} + \frac{f_{\text{bdy}}}{n_{\text{bdy}}} \times (n_{\text{cut}} - n_{\text{abv}})}{n_{\text{cut}}},$$

where fmr denotes false-match rate, f_{abv} is the number of false matches and n_{abv} the number of declared matches with weights above the cutoff weight, f_{bdy} is the number of false matches and n_{bdy} the number of declared matches with weights equal to the boundary cutoff weight, and n_{cut} is the number of declared matches needed to satisfy the condition that a certain percentage of the PES data file be declared matched. If we were to calculate the false-match rate by randomly selecting the appropriate number of boundary records to satisfy the cutoff criterion, then the expression above would give the expected false-match rate over repetitions of such a procedure; thus, the logic behind this definition is clear.

In the example above, one fourth of the boundary cases are false matches, and twenty additional records are needed to satisfy the stipulation that 80% of the file be declared matched. Effectively five false matches are added to the three among the records among the pairs with weights above the cutoff weight, giving a false-match rate of $(3 + 0.25(40 - 20))/8{,}000 = 8/8{,}000 = 0.001$.

3.7 Further Considerations Relevant to the Analysis of Experimental Results

Analysis of the experimental results proceeded from the standpoint that general indications of significance are more important than precise p-values, especially because the experiment itself is exploratory. Belin (1991) points out that appropriate methods for assessing significance from these data are somewhat complicated; this is because site should be thought of as a random factor (since we would like to generalize about treatment effects from the sample of three sites to a population of many possible sites), but standard procedures that use the site by treatment interaction as the error term for a particular treatment suffer from low power given the small number of available sites. Belin (1991) uses the Johnson-Tukey display-ratio plot (Johnson and Tukey 1987), which is a close relative of the half-normal plot of Daniel (1959), to estimate underlying noise levels in assessing the significance of effects. In this paper, we do not attempt to present formal significance findings.

4. RESULTS

4.1 ANOVA Breakdown of Experimental Results

We begin by breaking down the results of the factorial experiment into an analysis of variance, distinguishing treatment effects, site effects, cutoff effects, and their interactions from one another, grouping effects of the same order. Table 4.1 is an excerpt from the complete ANOVA breakdown of the experiment, showing treatment interactions up to four-way along with corresponding error terms.

F-statistics are calculated dividing the mean square for the given effect by the mean square for the effect-by-site interaction term. Thus, for example, the F-statistic for three-way interactions among treatments is calculated as $0.0120/0.00470 = 2.551$, with the denominator coming from the line for the four-way treatment-by-site interaction.

If the F-statistics are interpreted in the usual way, then statistical significance at the 0.0001-level is achieved for all of the F-statistics reported in Table 4.1 except the treatment-by cutoff four-way interactions; however, caution should be used in interpreting these results. First, the magnitudes of the various mean-square terms suggest that the higher-order effects are not of substantial practical importance. Further, the comparison of the F-statistics calculated above to a reference F-distribution relies on certain exchangeability assumptions (*e.g.*, that site-to-site variability in main effects is the same for all main effects) that are not necessarily well-founded. For example, it may not make sense to pool site-to-site variability in the effect of four versus seven digits of phone number with site-to-site variability in the effect of the different weighting schemes in estimating an error term for main effects.

4.2 Importance of Choice of Cutoff as Compared to Other Controllable Factors

It is evident (*e.g.*, from the mean squares for main effects) that site-to-site variability and variability due to the choice of a cutoff are considerably larger than the variability explained by differences in treatments. Although

Table 4.1

Excerpt from ANOVA Breakdown of Factorial Experiment, Grouping Effects of the Same Order

Source	*df*	Sums of squares	Mean square	*F*
Site main effects	2	35.195	17.598	
Treatment main effects	13	30.917	2.378	10.570
Cutoff main effects	12	147.515	12.293	7.548
Treatment/site 2-way interactions	26	5.850	0.225	
Cutoff/site 2-way interactions	24	39.089	1.629	
Treatment/treatment 2-way ints	70	6.992	0.100	4.041
Treatment/cutoff 2-way ints	156	1.410	0.009	3.553
Treatment/site 3-way interactions	140	3.461	0.0247	
Cutoff/treatment/site 3-way ints	312	0.794	0.0025	
Treatment 3-way interactions	206	2.472	0.0120	2.551
Treatment/cutoff 3-way ints	840	0.530	0.0006	1.866
Treatment/site 4-way interactions	412	1.938	0.00470	
Cutoff/treatment/site 4-way ints	1,680	0.568	0.00034	
Treatment 4-way interactions	365	0.747	0.00205	2.365
Treatment/cutoff 4-way ints	2,472	0.267	0.00011	0.236
Treatment/site 5-way interactions	730	0.632	0.00087	
Cutoff/treatment/site 5-way ints	4,944	0.226	0.00046	
. . .				
Total	**56,159**	**279.169**		

this result may be explained in part by the fact that some treatments are very close to one another (*e.g.*, using four digits versus seven digits of phone number), it is nevertheless the case that some of the qualitative differences between treatments are quite substantial (*e.g.*, leaving out two matching variables versus keeping them in). The ANOVA breakdown also highlights the fact that we can expect substantial site-to-site variability in false-match rates. In their approach to calibrating record-linkage procedures, Belin (1991) and Belin and Rubin (1991) explicitly accommodate site-to-site variability in providing estimates of false-match rates corresponding to different cutoffs.

4.3 The Main Effects of Treatments

In Table 4.2, we give the mean of the outcome variable observed for each level of the treatment factors. Since arcsine (x) is a monotone increasing function of x, lower values of the outcome signify lower false-match rates and thus better performance.

Table 4.2

Marginal Values of arcsine($\sqrt{\text{fmr}}$) for each Level of Experimental Treatments Averaged over all other Experimental Conditions

Factor	A	(name wts)	Factor	B	(inexact agree, name wts)	Factor	C	(non-name wts)
Level	1	0.106	Level	1	0.113	Level	1	0.101
	2	0.096		2	0.094		2	0.101
	3	0.093		3	0.095			
	4	0.130						
	5	0.079						
Factor	D	(inexact agree, non-name wts)	Factor	E	(Standardize name)	Factor	F	(Adjust for correlated agree)
Level	1	0.111	Level	1	0.102	Level	1	0.106
	2	0.108		2	0.100		2	0.095
	3	0.084						
Factor	G	(Include marit/rel)	Factor	H	(Four or seven digits phone #			
Level	1	0.103	Level	1	0.102			
	2	0.098		2	0.100			

Belin (1991) breaks down the experimental findings into a set of complementary orthogonal contrasts. The largest main-effect contrasts among those prespecified by Belin (1991) were those between frequency name weights (A = 5) and Fellegi-Sunter name weights (A = 4), between Winkler's string comparators on non-name fields (D = 3) and Jaro's corresponding string comparators (D = 2), between some string comparator for names (B = 2 or 3) and no string comparator for names (B = 1), between some string comparator for non-name fields (D = 2 or 3) and no string comparator for these fields (D = 1), and between performing an adjustment for correlated agreement (F = 2) and not performing such an adjustment (F = 1).

4.4 Two-Way Treatment Interactions

The largest two-way treatment interaction contrast among those reviewed by Belin (1991) was the F × G effect, which is the interaction of performing an adjustment for correlated agreement (among first name and middle initial and among first name, age, and sex) with including or not including marital status and relationship to head of household as matching variables. This contrast was statistically significant according to any of the procedures used in Belin (1991) for estimating a background noise level. We show the average levels of the outcome across the four treatment combinations above in Table 4.3.

Table 4.3
Average Performance for Combinations of F and G Treatments

F	G	False-match rate	Arcsine($\sqrt{\text{fmr}}$)
1	1	0.0182	0.116
1	2	0.0143	0.097
2	1	0.0128	0.091
2	2	0.0151	0.100

This result suggests that the adjustment for correlated agreement (level 2 of factor F) helps a great deal when marital status and relationship are not included as matching variables (level 1 of factor G), but the adjustment for correlated agreement does not help on average when marital status and relationship are included as matching variables. That we are able to identify this type of effect emphasizes the importance of pursuing empirical evaluations in an experimental framework.

The next two largest two-way treatment interaction contrasts cited by Belin (1991) after the F × G interaction comprise part of the A × B interaction (involving the choice of name weights and the choice of string comparisons to use for name fields). We show the average results for all of the combinations of treatments for factors A and B below as Table 4.4.

Table 4.4
Average Performance for Combinations of A and B Treatments

A	B	False-match rate	Arcsine ($\sqrt{\text{fmr}}$)
1	1	0.0192	0.120
1	2	0.0140	0.099
1	3	0.0143	0.100
2	1	0.0170	0.110
2	2	0.0120	0.087
2	3	0.0123	0.089
3	1	0.0177	0.113
3	2	0.0118	0.084
3	3	0.0119	0.083
4	1	0.0254	0.145
4	2	0.0193	0.123
4	3	0.0189	0.122
5	1	0.0109	0.079
5	2	0.0109	0.079
5	3	0.0109	0.078

Thus, we find that when we use frequency-based name weights (A = 5), it hardly matters whether we use any string comparison method, but when we use *ad hoc* name weights or Fellegi-Sunter name weights, the use of string comparison methods substantially improves the average performance of the computer-matching procedure.

We highlight some of the other interesting findings noted in Belin (1991) based on exploring the largest two-way treatment interaction effects:

(1) The Winkler approach to inexact agreement on non-name variables (*i.e.*, D = 3), which is the best treatment on average for factor D, has more of a helpful effect on average when marital status and relationship to head of household are included as matching variables (*i.e.*, G = 2), even though the latter variables are not included in any of the treatments for handling inexact agreement.

(2) Unlike the other treatments for name weights, which appear to be helped by the inclusion of marital status and relationship, frequency-based name weighting appears to be adversely affected by the inclusion of these variables.

(3) *Ad hoc* weights of ±6 for agreement on name perform better on average when combined with the *ad hoc* weighting approach to non-name variables; *ad hoc* name weights of ±4 and ±2 work better with the weights assigned by the Fellegi-Sunter algorithm to non-name variables.

(4) Without the adjustment for correlated agreement, Fellegi-Sunter weights for non-name variables worked better for these data than *ad hoc* weights, but the *ad hoc* weights worked better when the adjustment for correlated agreement was included. (However, based on the method of estimating the background noise level described in Belin (1991), this phenomenon should not necessarily be expected to carry over to other sites.)

4.5 Which Treatment Combination Works Best?

To wrap up the analysis of the experimental results, we consider now the question of which treatment combination works best. To measure the performance for a given treatment combination, we take the average outcome from using that procedure across the three available sites. The outcomes we examine are the false-match rates corresponding to 60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5%, and 90% of the PES file declared matched. The results from the experiment are provided in Table 4.5.

Table 4.5

Best Treatment Combination for each of Thirteen Cutoffs from Factorial Experiment

Cutoff level	Levels of factors in best treatment combination (A B C D E F G H)	False-match rate for best treatment combination averaged over three sites
60% matched	3 3 2 3 2 1 1 1	0.00042
62.5% matched	3 3 1 3 1 1 2 1	0.00047
65% matched	3 3 1 3 2 2 2 2	0.00052
67.5% matched	3 3 2 3 2 2 1 1	0.00071
70% matched	2 3 2 3 1 2 1 2	0.00079
72.5% matched	5 2 2 3 1 1 1 2	0.00081
75% matched	5 1 1 3 2 1 2 1	0.00112
77.5% matched	3 3 1 3 2 1 2 1	0.00133
80% matched	2 3 2 3 1 2 1 2	0.00188
82.5% matched	3 3 1 3 2 2 1 1	0.00571
85% matched	5 1 2 3 2 2 1 2	0.01556
87.5% matched	2 3 2 3 1 2 1 2	0.03023
90% matched	2 3 2 3 1 2 1 2	0.05174

These results contrast with the earlier result suggesting that frequency-based weighting for names (level 5 for factor A) is better on average than using *ad hoc* name weights of ± 6 (level 3 for factor A). Apparently, the reason that the latter is worse on average is due to certain interaction effects. When the *ad hoc* weighting approach is combined with the appropriate levels of other factors, it appears to perform at least as well as the frequency-weighting approach. We also note that the best combination of factors F and G is not always treatments 2 and 1, respectively, despite our earlier finding that this treatment combination for these two factors performs best on average. Only treatment 3 of factor D (using Winkler modifications in handling inexact agreement on non-name variables) is an unequivocal choice for the best treatment no matter how we measure the outcome of the experiment. The choice for the best treatment for name weights is between deterministic weights of ± 6 or ± 4 and the frequency name-weighting approach. If one of the deterministic weighting schemes is used, the Winkler approach to string comparisons for names is to be recommended; with frequency name weights, it is not clear that any string comparison approach should be used on names.

Between Fellegi-Sunter weights for non-name variables and *ad hoc* weights, the choice is not obvious, but earlier analysis suggested that the effect either way is small. Similar remarks apply to the choice of whether to use standardized or unstandardized first names and to the choice of whether to use four or seven digits of the phone number.

Considering the fact that there is not a single treatment combination that is uniformly superior to all other treatment combinations, one might look to the performance of different treatment combinations in a particular region of interest (*e.g.*, where the false-match rate is around 0.001). However, if we look at the best treatment combinations in the region where 70%-80% of the PES file is declared matched (*i.e.*, restricting attention to five cutoffs), we still find no obvious choice for a preferred treatment combination. Averaged across those five cutoffs, the best treatment combination is (2,3,2,3,1,2,1,2); that is, using name weights of ± 4, incorporating Winkler's modifications to inexact agreement on name, estimating weights using the Fellegi-Sunter algorithm for non-name variables, using Winkler's approach to inexact agreement for non-name variables, using the original unstandardized version of first name, adjusting the composite weight for correlated agreement, not including marital status and relationship to head of household as matching variables, and using all seven digits of phone number.

For comparison, we display in Table 4.6 the average performance of some of the other candidates for best treatment combination. Thus it appears that the best alternatives to (2,3,2,3,1,2,1,2) are treatment combinations (3,3,1,3,2,2,2,2) and (3,3,1,3,2,1,2,1). Both of these procedures feature name weights of ± 6, predetermined

Table 4.6

Average False-match Rates for Different Treatment Combinations Across Three Sites and across Five Cutoff Levels (70%, 72.5%, 75%, 77.5%, and 80% of PES File Declared Matched)

Levels of factors in treatment combination (A B C D E F G H)	Average false-match rate across sites and across cutoffs with 70%, 72.5%, 75%, 77.5%, and 80% of PES file declared matched
3 3 2 3 2 1 1 1	0.00493
3 3 1 3 1 1 2 1	0.00154
3 3 1 3 2 2 2 2	0.00137
3 3 2 3 2 2 1 1	0.00161
2 3 2 3 1 2 1 2	0.00124
5 2 2 3 1 1 1 2	0.00191
5 1 1 3 2 1 2 1	0.00153
3 3 1 3 2 1 2 1	0.00138
3 3 1 3 2 2 1 1	0.00156
5 1 2 3 2 2 1 2	0.00155

ad hoc weights for non-name variables, Winkler's approaches to inexact agreement for both name and non-name variables, standardized first names, and inclusion of marital status and relationship as matching variables. These treatment combinations differ from each other in that one includes an adjustment of the composite weight for correlated agreement and calls for using seven digits of phone number, whereas the other features no adjustment of weights for correlated agreement and only four digits of phone number. The treatment combinations involving the use of frequency-based name weighting do not perform as well as the best treatment combinations using *ad hoc* name weights according to this standard.

In the 1990 PES, the treatment combination that was used in computer-matching operations was very close to treatment combination (5,3,2,3,2,2,2,1). In the test-census data sets studied here, this treatment combination produced an average false-match rate across the five cutoffs of 0.00179.

4.6 Concluding Remarks

While the results in this paper address the tradeoff between the number of records declared matched and false-match rates, an anonymous referee noted that "every gain which is achieved by a superior record linkage procedure must be justified by the cost of implementing that procedure." This is another tradeoff that any practitioner can appreciate. Hopefully, the findings presented here about the relative importance of various factors in record linkage will provide some guidance to those who develop and implement linkage software. Because some of the results may depend on specific features of the census/PES data being matched, there may be some question as to how these results relate to other record-linkage settings. But as was emphasized at the outset, one practical recommendation that does generalize across data settings is the call for taking an experimental approach to the study of record linkage. Empirical study through designed experiments is a tried and true source of guidance, offering a clear framework for adding to the accumulated insights of record-linkage specialists.

ACKNOWLEDGMENTS

Much of this work was done while the author was working for the Record Linkage Staff of the U.S. Bureau of the Census in Washington, D.C. The author gratefully acknowledges helpful discussions and comments from Don Rubin, Bill Winkler, Alan Zaslavsky, and an anonymous referee, as well as earlier support from JSA 88-02 and JSA 89-07 while the author was a doctoral candidate at Harvard University.

REFERENCES

ABBATT, J.D. (1986). A cohort study of eldorado uranium workers. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 51-57.

ACHESON, E.D. (1967). *Medical Record Linkage*. Oxford: Oxford University Press.

ACHESON, E.D. (Ed.) (1968). *Record Linkage in Medicine*, Edinburgh: E. & S. Livingstone.

BALDWIN, J.A., ACHESON, E.D., and GRAHAM, W.J. (Eds.) (1987). *A Textbook of Medical Record Linkage*. Oxford: Oxford University Press.

BELIN, T.R. (1989a). Outline of procedure for evaluating computer matching in a factorial experiment. Unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BELIN, T.R. (1989b). Results from evaluation of computer matching. Unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BELIN, T.R. (1991). Using mixture models to calibrate error rates in record-linkage procedures, with application to computer-matching for census undercount estimation. Ph.D. thesis, Department of Statistics, Harvard University. (Published by University Microfilms, Inc.)

BELIN, T.R., and RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C.

BOYLE, C.A., and DECOUFLÉ, P. (1990). National sources of vital status information: Extent of coverage and possible selectivity in reporting. *American Journal of Epidemiology*, 131, 160-168.

BROWN, P., LAPLANT, W., LYNCH, M., ODELL, S., THIBAUDEAU, Y., and WINKLER, W. (1988). Collective Documentation for the 1988 PES Computer Match Processing and Printing. Vols. I-III, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BUREAU OF THE CENSUS (1988-1991). 1990 Decennial Census Information Memorandum Series, Decennial Planning Division, Bureau of the Census, Washington, D.C.

[Note: To all of the reports in the aforementioned memorandum series, the following statement is attached:

"These overviews are prepared for use by planning and operating divisions within the Census Bureau who are conversant with the background, previous experiences, terminology, and processes, as well as with the overall framework of the decennial census design, goals, and inter-relationships of operations and systems. They are NOT [emphasis in original] intended or appropriate for external distribution and should not be sent outside the Census Bureau without prior approval from Jim Dinwiddie ([301]-763-5270) of the Decennial Planning Division."]

BUREAU OF THE CENSUS (1987-1991). STSD Decennial Census Memorandum Series, Statistical Support Division, U.S. Bureau of the Census, Washington, D.C.

CARPENTER, M., and FAIR, M.E. (Eds.) (1990). Canadian Epidemiology Research Conference - 1989: *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Centre for Health Information, Statistics Canada, Ottawa, Ontario.

CHERNOFF, H. (1980). The identification of an element of a Large population in the presence of noise. *Annals of Statistics*, 8, 1179-1197.

CHILDERS, D. (1989). 1990 PES Within Block Matching - Clerical Matching Group. STSD Decennial Census Memorandum Series #V-69, U.S. Bureau of the Census, Washington, D.C.

CITRO, C.F., and COHEN, M.L. (Eds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, Washington, D.C.: National Academy Press.

COHEN, M.L. (1990). Adjustment and reapportionment - Analyzing the 1980 decision. *Journal of Official Statistics*, 6, 241-250.

COOMBS, J.W., and SINGH, M.P. (Eds.) (1988). *Proceedings of the Symposium on Statistical Uses of Administrative Data*. Statistics Canada, Ottawa, Ontario.

COPAS, J., and HILTON, F. (1990). Record Linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society*, Series A, 153, 287-320.

CURB, J.D., FORD, C.E., PRESSEL, S., PALMER, M., BABCOCK, C., and HAWKINS, C.M. (1985). Ascertainment of vital status through the National Death Index and the Social Security Administration. *American Journal of Epidemiology*, 121, 754-766.

DANIEL, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1, 311-341.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

DONOGHUE, G. (1990). Clerical Specifications for the 1990 Post Enumeration Survey Before Followup Matching - Special Matching Group. STSD Decennial Census Memorandum Series #V-92, U.S. Bureau of the Census, Washington, D.C.

DULBERG, C.S., SPASOFF, R.A., and RAMAN, S. (1986). Reactor clean-up and bomb test exposure study. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 59-62.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and Beyond (with discussion). *Journal of the American Statistical Association*, 80, 98-131.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association*, 84, 927-944.

FAGERLIND, I. (1975). *Formal Education and Adult Earnings: A Longitudinal Study on the Economic Benefits of Education*, Stockholm: Almquist and Wiksell.

FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (with discussion). *Statistical Science*, 1, 1-39.

GOLDACRE, M.J. (1986). The Oxford record linkage study: Current position and future prospects. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press, 106-129.

HILL, T. (1981). Generalized Iterative Record Linkage System: GIRLS. (Glossary, Concepts, Strategy Guide, User Guide), Systems Development Division, Statistics Canada, Ottawa, Ontario.

HILL, T., and PRING-MILL, F. (1986). Generalized iterative record linkage system: GIRLS, (revised edition). Systems Development Division, Statistics Canada, Ottawa, Ontario.

HOGAN, H. (1992). The 1990 Post-Enumeration Survey: An Overview. *The American Statistician*, 46, 261-269.

HOWE, G.R., and LINDSAY, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers in Biomedical Research*, 14, 327-340.

HOWE, G.R., and SPASOFF, R.A. (Eds.) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press.

JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: Methodological issues. *Journal of Official Statistics*, 2, 255-277.

JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

JOHANSEN, H.L. (1986). Record linkage of national surveys: The Nutrition Canada example. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 153-163.

JOHNSON, E.G., and TUKEY, J.W. (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao Data. In *Design, Data, and Analysis*, (Ed. C.L. Mallows) New York: John Wiley and Sons.

JOHNSON, R.A. (1991). Methodology for Evaluating Errors in U.S. Department of Justice Attorney Workload Data. Unpublished technical report, General Accounting Office, Washington, D.C.

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.

KERSHAW, D., and FAIR, J. (1979). *The New Jersey Income and Maintenance Experiment: Operations, Surveys, and Administration*, Volume I. New York: Academic Press.

KILSS, B., and ALVEY, W. (Eds.) (1984a). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. I, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and ALVEY, W. (Eds.) (1984b). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. II, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and ALVEY, W. (Eds.) (1984c). *Statistics of Income and Related Administrative Record Research:* 1984, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and ALVEY, W. (Eds.) (1985). *Record Linkage Techniques - 1985*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and ALVEY, W. (Eds.) (1987). *Statistics of Income and Related Administrative Record Research: 1986-1987*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and JAMERSON, B. (Eds.) (1990). *Statistics of Income and Related Administrative Record Research 1988-1989*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and SCHEUREN, F. (1978). The 1973 CPS-IRS-SSA Exact Match Study. *Social Security Bulletin*, Vol. 41, 10, 14-22.

LAPLANT, W. (1988). User's Guide for the Generalized Record Linkage Program Generator (GENLINK). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

LAPLANT, W. (1989). User's Guide for the Generalized Address Standardizer (GENSTAN). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.

NEWCOMBE, H.B., FAIR, M.E., and LALONDE, P. (1992). The use of names for linking personal records (with discussion). *Journal of the American Statistical Association*, 87, 1193-1208.

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., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., and ABBATT, J.D. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado Uranium Workers. *Computers in Biology and Medicine*, 13, 157-169.

NICHOLL, J.P. (1986). The use of hospital in-patient data in the analysis of the injuries sustained by road accident casualties. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 243-244.

PALETZ, D. (1989). Name standardization software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

ROGOT, E., SORLIE, P.D., and JOHNSON, N.J. (1986). Probabilistic methods in matching census samples to the National Death Index. *Journal of Chronic Disease*, 39, 719-734.

ROGOT, E., SORLIE, P.D., JOHNSON, N.J., GLOVER, C.S., and TREASURE, D.W. (1988). A Mortality Study of One Million Persons. Public Health Service, National Institutes of Health, Washington, D.C.

SCHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions (with discussion). *Journal of the American Statistical Association*, 82, 965-990.

SMITH, M.E., and NEWCOMBE, H.B. (1975). Methods for computer linkage of hospital admission-separation records for cumulative health histories. *Methods of Information in Medicine*, 14, 118-125.

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

THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Section on Statistical Computing, American Statistical Association* 283-288.

WENTWORTH, D.N., NEATON, J.D., and RASMUSSEN, W.L. (1983). An evaluation of the Social Security Administration Master Beneficiary Record File and the National Death Index in the ascertainment of vital status. *American Journal of Public Health*, 73, 1270-1274.

WILLIAMS, B.C., DEMITRACK, L.B., and FRIES, B.E. (1992). The accuracy of the National Death Index when personal identifiers other than Social Security Number are used. *American Journal of Public Health*, 82, 1145-1147.

WINKLER, W.E. (1985a). Preprocessing of lists and string comparison. In *Record Linkage Techniques - 1985*, (Eds. W. Alvey and B. Kilss). 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*, (Eds. W. Alvey and B. Kilss). Internal Revenue Service, Publication 1299 (2-86), 227-241.

WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter Model of record linkage. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 145-155.

WINKLER, W.E. (1991). Documentation of record-linkage software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

WINKLER, W.E., and THIBAUDEAU, Y. (1992). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.