Journal of Official Statistics
Vol. 2. No. 3, 1986, pp. 255–277
© Statistics Sweden

Record Linkages for Statistical Purposes: Methodological Issues

Thomas B. Jabine[1] and Fritz J. Scheuren[2]

Abstract: In spite of their relative complexity, model-based computer record-linkage systems have many advantages over ad hoc procedures. This article begins with a brief review of the underlying theory and models, developed in the 1950s and '60s, for such record-linkage systems. The authors then identify and discuss key decisions that are necessary in developing a record-linkage system or using an existing system. For each of the three main stages of record linkage – preprocessing, matching, and analysis – current practices and unsolved problems are reviewed. Recommendations for research and development are presented. The article draws in part on papers presented at a May 1985 Workshop on Exact Matching Methodologies, held in Arlington, VA, U.S.A. with participants from the United States and Canada.

Key words: Matching; record-linkage theory and models; file standardization; validation; blocking; weights; estimation.

1. Introduction

Records from different data files or sources are linked for many purposes. Statistical purposes include development of frames for censuses and surveys, enhancement of survey data by adding data from administrative files and the evaluation of coverage and response errors in censuses and surveys. Record linkage is a key element in the dual systems method of estimating vital rates, first proposed by Chandrasekaran and Deming (1949), and in the multiple system estimation methods developed since then (Marks et al. (1974)).

Nonstatistical purposes for record linkages include the development and maintenance of record systems for such widely varying activities as tax administration and compliance, public and private insurance, banking and credit, and criminal justice, to mention but a few. The focus of this paper is on statistical applications; however, much of what will be presented applies equally well to nonstatistical applications.

1.1. Historical observations

Prior to the advent of computers, most record linkages were carried out under ad hoc rules and procedures. Human judgment played a major role. Borderline cases were often resolved, not by recourse to explicit rules, but by giving the cases to one or more "experts" deemed best qualified to judge whether or not two records were associated with the same entity.

[1] Statistical consultant, 3231 Worthington Street, N.W., Washington, D.C. 20015, USA.

[2] Director, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Avenue, N.W., Washington, D.C. 20224, USA.

As computers began to be widely available in the 1950s, it was natural to try to exploit their power and reliability for large-scale record-linkage activities. Like other applications of computers, this required a careful review and rethinking of procedures and the development of explicit specifications for every step in the process.

The main theoretical underpinnings for computer-oriented matching methods were firmly established by the late 1960s with the papers of Tepping (1968) and Fellegi and Sunter (1969). Sound practice dates back even earlier, at least to the 1950s and the work of Newcombe and his collaborators (e.g., Newcombe et al. (1959)).

Despite these promising early developments, most record linkages in the 1970s continued to be performed with ad hoc heuristic methods. There were many reasons for this:

- First, until recently (and maybe even now) there have been only a few people whose main professional interest is data linkage. This means, among other things, that most of the applied work in this field has been done by individuals who may be solving matching problems for the first time. Because the basic principles of matching are deceptively simple, some ad hoc solutions have probably been far from optimal.
- Second, statisticians typically get involved very late in the matching step, often after the files have already been matched and a new file created. Even when this is not the case, little emphasis may be placed on the data structures needed for linkage, because other uses of the data are assigned higher priorities. Design opportunities have, therefore, been generally limited to what steps to take given that the files were produced largely for other purposes.
- Third, until the late 1970s good, portable, general-purpose matching software had not been widely available (e.g., Howe and Lindsay (1981)), despite some important early attempts (e.g., Jaro (1972)). Even in the presence of general-purpose software, the uniqueness of each matching environment may lead practitioners to write complex customized programs, thereby absorbing resources that might have been better spent in other ways.
- Fourth, especially for matches to administrative records, barriers to the introduction of improved methods have existed because cruder methods were thought to be more than adequate for administrative purposes.
- Fifth, the analysis of linked data sets, with due consideration to matching errors, is still in its infancy. Qualitative statements about such limitations typically have been all that practitioners have attempted.

More will be said below concerning these issues in the context of computerized matching.

1.2. Recent developments

The 1980s brought a resurgence of interest in the development of sophisticated general-purpose computerized systems for record linkage, based on models similar to those proposed by the pioneers of the 1950s and 1960s. Important new model-based systems were developed to link exposure and mortality records in epidemiological followup studies: The Generalized Iterative Record Linkage System (GIRLS) developed by Statistics Canada and the National Cancer Institute of Canada (Smith and Silins (1981)) and the California Automated Mortality Linkage System (CAMLIS) developed by the State of California (Arellano (1985)). The U.S. Bureau of the Census began work on a generalized record-linkage system, designed primarily to evaluate census coverage by linking census and administrative records (Jaro (1985)). Two U.S. agencies that conduct economic surveys, the Statistical Reporting Service of the Department of Agriculture (Coulter (1985)) and the Energy Information Administration (Winkler

(1985a)), have developed record-linkage systems for use in constructing sampling frames from multiple list sources. At the start of this decade the National Center for Health Statistics established the National Death Index (NDI), which contains computerized records for all deaths occurring in the United States from 1979 on. Health and medical researchers use the NDI to determine which persons in their study populations have died. Research and operating experience have led gradually to improvements in the record-linkage procedures used in the NDI operations (Patterson and Bilgrad (1985)).

In May 1985 the Washington Statistical Society (a local chapter of the American Statistical Association) and the U.S. Federal Committee on Statistical Methodology sponsored the Workshop on Exact Matching Methodologies. The workshop, with an attendance of 140 persons from the United States and Canada, brought together the architects of the record-linkage models and systems mentioned above and many other persons engaged in or interested in record-linkage theory and applications.

The presentations and discussions at the workshop provided valuable information on current theory and practice and included suggestions for future research and development. The proceedings of the workshop (U.S. Internal Revenue Service (1985)) include the papers presented, the discussants' comments and the recommendations of the interagency committee that organized the workshop.[3]

1.3. Purposes and organization of paper

The purpose of this paper is to review briefly the historical development of record-linkage theory and applications, to identify and discuss matching design questions that must be looked at in developing record-linkage systems or specific applications and to suggest some questions for further study. The authors have drawn heavily on the presentations at the May 1985 Workshop on Exact Matching Methodologies.

For this paper, the term record linkage is intended to cover only exact matching, i.e., "a match in which the linkage of data for the same unit (e.g., person) from the different files is sought" (U.S. Office of Federal Statistical Policy and Standards (1980)). As pointed out earlier, although the focus is on statistical applications, many of the issues discussed are just as relevant to nonstatistical applications. However, the consequences of matching errors are likely to be quite different for statistical and nonstatistical applications.

Much of the discussion is also general with respect to the kinds of data files and records to be linked. The records may be for persons or for other units, such as households, business establishments or companies. Some of the input files may have been developed initially for statistical purposes, others for administrative purposes. A file may represent an entire defined population or only a sample from a defined population. Files to be linked may have the same or different time references. In the most general context for record linkages, input files may contain duplicates, so that linkages may be attempted both within and between files.

Section 2 of this paper briefly discusses the theoretical framework for record linkage, with emphasis on the Fellegi-Sunter model. Section 3 identifies the steps in a record-linkage application and presents some general design issues that must be considered in the planning stage. Sections 4, 5 and 6 review current practice and unsolved problems in the three main stages of record linkage: preprocessing, matching and analysis. Section 7 presents our

[3] A copy of the proceedings, *Record Linkage Techniques – 1985*, may be obtained by writing: Statistics of Income Division, Internal Revenue Service D:R:S, 1111 Constitution Avenue, N.W., Washington, D.C. 20224, USA.

conclusions, with emphasis on recommendations for future research and development.

The authors favor the use of model-based methods of record linkage over ad hoc methods. Tepping gives a concrete example in which the use of his model reduced expected costs (including the costs associated with errors in matching) by about 50 percent in comparison with an ad hoc procedure. As Tepping points out, " The parameters [of the model] may be difficult to determine. Also, it will be seen, the mathematical model (as usual) is not an exact representation of the real world. Nevertheless, the model provides useful guides for the construction of efficient linkage rules..." (Tepping (1968, p. 1322)).

2. Record Linkage Theory

This section begins with a brief summary and discussion of the Fellegi-Sunter model. Contributions of others, especially Tepping, are then reviewed.

The Fellegi-Sunter model starts with two files of records, A and B. The object of the record linkage is to recognize the records in the two files which represent identical persons (or other kinds of units). All possible pairs of records, one from each file, are to be examined. These pairs of records (a, b) are called comparison pairs. To be a match, the comparison pair (a, b) must consist of records for the same entity. Accordingly, a pair consisting of records for two different entities is a nonmatch or an unmatched pair.

For each comparison pair, one of three linkage decisions, $D(i)$ for $i = 1,2,3$, is to be made. The decisions are:

$D(1)$ – a and b are for the same unit (called a positive link).

$D(2)$ – a decision is not possible without further investigation (called a possible link).

$D(3)$ – a and b are for different units (called a positive nonlink).

The linkage decision taken for each comparison pair depends on rules based on the extent of agreement observed between the values of the matching variables for records a and b. In the Fellegi-Sunter model, the decisions are based on ratios of conditional probabilities: the probability of the observed result of the comparison, given that a and b are in fact for the same unit and the probability of the same observed result, given that a and b represent different units.

Decisions may, of course, be incorrect. A false match occurs when the decision $D(1)$ is made and a and b represent different units. A false nonmatch occurs when the decision $D(3)$ is taken and a and b represent the same unit. For the Fellegi-Sunter model, the respective probabilities of false matches and false nonmatches are denoted by μ and λ. Different decisions have different costs. Operational costs are likely to be highest for the decision $D(2)$, which specifies further investigation to make a positive determination.

A linkage rule associates one of the three decisions $D(i)$ with every possible result of observing the values of the matching variables for a pair of records (a, b). In the Fellegi-Sunter model, an optimum linkage rule is one which achieves specified values of μ and λ and minimizes the number of pairs classified as possible links (decision $D(2)$).

As Fellegi and Sunter (1969) point out, their theory for record linkage could have been formulated in terms of the classical theory of hypothesis testing, with their form of linkage rule being equivalent to a likelihood ratio test and their optimum linkage rules being the uniformly most powerful test for the alternative null hypotheses of the pair (a, b) being a nonmatch or a match. Kirkendall (1985) has shown that the test statistic and optimum linkage rule used in their model can also be derived by using an information theoretic approach (Kullback (1968)).

While their basic model calls for the comparison of all possible pairs (a, b) formed by elements of files A and B, Fellegi and Sunter also consider the possibility of blocking, i.e., restricting the comparisons to pairs for which a and b are in agreement for one or more matching variables. Given the current power of computers, examination of all possible pairs (a, b) is seldom economically feasible, even for medium-size files, so some form of blocking is usually employed. The most likely effect of blocking is to decrease the probability of false matches and increase the probability of false nonmatches. The first of these outcomes is, of course, desirable; the second is not. Fellegi and Sunter examine these effects and discuss methods of choosing among alternative blocking procedures. Kelley (1984, 1985) provides further guidance on how to make an objective choice among alternative blocking procedures by weighing the reduced costs of computation against the errors introduced by not looking at all comparison pairs.

As indicated earlier in the citation from Tepping, one of the key questions in using model-based record-linkage systems is how to estimate the parameters of the model. For the Fellegi-Sunter model, the problem is to estimate the likelihood ratios, often called weights, associated with the possible outcomes of the comparisons. Fellegi and Sunter describe two methods for estimating weights. The first method assumes the availability of prior information on the distribution of the matching variables in the populations from which files A and B are drawn, as well as on the probabilities of errors in generating the individual records that are compared. An illustration of the method is given in Section 5 of this paper. The second method, which requires an assumption that errors in recording different matching variables be independent, estimates the components of the weights directly from the files being linked.

Under the Fellegi-Sunter model, each comparison pair is examined and classified into one of the three decision categories independently of all other pairs. Consequently, a record in file A can be classified as a positive link with two or more records in file B and vice versa. This may or may not be a satisfactory outcome, depending on the objectives of the record linkage and on what is known or assumed about the possibility of duplication in either of the input files. The appearance of groups of linked comparison pairs with common elements must be dealt with in practical record linkage applications. Howe and Lindsay (1981) and Kirkendall (1985) discuss methods that are appropriate, given various assumptions about duplication in the input files. In the matching system being developed by the Census Bureau, the initial assignment of positive links is done by an optimization process that does not permit any multiple linkages. Subsequently, however, the weights of other comparison pairs involving the linked records are systematically reviewed to identify possible duplicates (Jaro (1985)).

The Tepping (1968) model for record linkage uses the same underlying framework as the Fellegi-Sunter model, but differs in some important ways. The Fellegi-Sunter model permits only three possible decisions for each comparison pair: positive link, possible link and positive nonlink, and in some applications the possible link category is dispensed with (e.g., Howe and Lindsay (1981)). Tepping's model does not restrict the number of possible decisions to be taken for a comparison pair. He gives an example with five alternatives which could be characterized as positive link and nonlink, tentative link and nonlink, and possible link.

Fellegi and Sunter point out that costs, or losses associated with each of the possible decisions can be taken into account in setting the error levels μ and λ. Tepping, however, makes costs an explicit element of his model. Costs include both operational costs of record linkage and losses associated with matching errors. For each decision $D(i)$, cost is assumed to be a function of the conditional probability of a match, $P(\text{Match}|\gamma)$ where γ is the outcome of the comparison for a pair. The linkage rule is simple: for any specific value of $P(\text{Match}|\gamma)$, choose the decision $D(i)$ that has the smallest cost. This rule clearly minimizes the overall costs and is therefore an optimum linkage rule.

The parameters that must be estimated to apply the Tepping model differ from those needed for the Fellegi-Sunter model. The parameters include both the cost functions for the decisions $D(i)$ and the values of $P(\text{Match}|\gamma)$ for each possible comparison outcome. In order to assign the values of $P(\text{Match}|\gamma)$, one needs to estimate what proportion of the comparison pairs with each possible outcome are, in fact, matches. Tepping proposes that this be done by taking a sample of pairs for each outcome and attempting to determine their true match status, presumably on the basis of additional information obtained for these cases.

The two models have many elements in common and both provide useful guidelines for practical record-linkage operations. An important consideration in choosing between them would be a judgment of the feasibility of estimating the necessary parameters that, as we have seen, are different for the two models.

3. Applications: General Design Issues

Record linkages can be thought of as consisting of three phases: preprocessing, matching, and analysis.

(1) *Preprocessing* consists of all operations prior to the actual comparison of records. It includes the development or acquisition of input files and the operations, such as validation and standardization, performed on these files to facilitate the matching process. It also includes, in virtually all linkages, some form of blocking to limit the number of comparison pairs to be processed.

(2) The specific steps included in *matching* depend on the particular system being used. In probability-based systems, a weight is calculated for each comparison pair, based on the observed extent of agreement between the two records. The weights are the primary basis for decisions about the match categories to which the comparison pairs are assigned. Following the initial decisions on match status, additional steps may be needed, e.g., to reclassify "possible links" as positive links or nonlinks and to resolve multiple links.

(3) The *analysis* phase consists of all post-match activities, including the creation, use and dissemination of outputs and evaluation of the results.

These three phases are dealt with further in the sections which follow. Before discussing them in detail, however, it may be worthwhile to look at some key choices that should be made, at least tentatively, at the start of any record-linkage project: choice of input files and matching variables; choice of a record-linkage system; and specification of desired outputs. Early consideration of these aspects will usually pay significant dividends in quality and efficiency.

3.1. Choice of input files and matching variables

The choices of input files and matching variables are closely related. Success in linking records from two files will depend on the precise definitions used and the quality of reporting for variables included in both files. If the candidate files for a linkage already exist, potential matching variables must be evaluated to determine whether a linkage of acceptable quality is feasible.

If one or more of the input files has not yet been created, it may be possible to influence file development in ways that will facilitate record linkage, e.g., by adding variables needed for linkage to other files and by using operating definitions and formats that are compatible with those files. Kasprzyk (1983) describes procedures developed to maximize the completeness and accuracy of reporting of social security numbers in the U.S. Survey of Income and Program Participation (his purpose was to facilitate enhancement of the survey results through linkages with various kinds of administrative records).

To illustrate the issues involved in the choice of matching variables, let us consider the problem of linking U.S. files of person records based on common identifying information: social security number, name, address, sex, and birth date.

The social security number (SSN) is the most important linking variable that we in the United States have for person matching purposes. SSNs were first issued so that the earnings of persons in employment covered by the Social Security program could be reported for eventual use in determining benefits. Other uses by federal and state governments followed and now the SSN is a nearly universal identifier. It is also nearly a unique identifier all by itself and extremely well reported, both in survey settings and on records such as death certificates (e.g., Cobleigh and Alvey (1975), and Alvey and Aziz (1979)). In survey contexts, error rates may run to 2 or 3 percent; but this depends greatly on the extent respondents are required to use records to provide the requested information. Typically, drivers' licenses, pay stubs, and the like are excellent sources (in addition to the social security card itself).

Both administrative and survey reporting of social security numbers are subject to possible mistakes in processing, but these can be guarded against by using part of the individual's surname as a confirmatory variable. Both the Internal Revenue Service and the Social Security Administration use this method as one way of spotting keying errors. A difficulty with this approach is that name changes (especially for females) may lead to considerable extra effort in confirming (usually through correspondence) that the SSN was indeed correct to begin with.

One disadvantage of the SSN as an identifier and linking variable is the absence of an internal check digit allowing one to spot errors by a simple examination of the number itself (Scheuren and Herriot (1975)). A second problem is that people sometimes report another person's SSN as their own, usually unintentionally.

The extent to which this problem exists is unknown, but it is believed, at least by some authorities, to be less prevalent than the opposite problem – issuances of multiple numbers to the same person (U.S. Department of Health, Education and Welfare (1973)). Until 1972, applicants for SSNs were not asked if they had already been issued numbers, nor was proof of identity sought. This led to perhaps as many as 6 million or more individuals having two or more SSNs (Scheuren and Herriot (1975)). A substantial fraction of the multiple issuances have been cross-referenced so that multiple reports for the same individual can be brought together if desired. Based on work done as part of the 1973 Exact Match Study (Kilss and Scheuren (1978)), it appears that, despite the frequency of the problem, multiple issuances can largely be ignored unless one is looking at longitudinal information stretching back to the early days of the Social Security program.

On balance the SSN is nearly ideal as a linking variable, but it is not always available. For example, in the Current Population Survey the number is missing for adults between 20 and 30 percent of the time (Scheuren (1983)). However, evidence from work done in connection with the Survey of Income and Program Participation suggests that with a modest effort the SSN missed rate can be lowered significantly, to less than 10 percent, in Census Bureau surveys (Kasprzyk (1983)). Recent experience with death certificates shows a missed rate of about 6 percent for adults (Patterson and Bilgrad (1985)).

What can we do when the SSN is missing or proves unusable? We are obviously forced either to seek more information or to try to link records using the other matching variables. As a rule, none of these other variables is unique alone and all of them, of course, are subject in varying degrees to reporting problems of their own. Some examples of the problems typically encountered are:

- *Surname* – As already mentioned, name changes due to marriage or divorce are, perhaps, the main difficulty. For some ethnic groups, there can be many last names and the order of their use may vary.

- *Given Name* – The chief problem is the widespread use of nicknames. Some are readily identifiable ("Jim" for "James") but others are not (like "Stony" for "Paul").

- *Middle Initial* – People may have many middle names (including their maiden name) and the middle name they employ may vary from occasion to occasion. Often, too, this variable may be missing (Patterson and Bilgrad (1985)).

- *Address* – This is an excellent variable for confirming otherwise questionable links. Disagreements are hard to interpret, however, because of address changes; address variations (e.g., 21st and Pennsylvania Avenue for 2122 Pennsylvania Avenue); and, of course, differences between mailing addresses (usually all that is available in administrative files) and physical addresses (generally all that is obtained in a household survey). Recent research on this variable has been done by Childers and Hogan (1984).

- *Sex* – Sex is generally well reported and, except for processing errors, can be relied upon. The main difficulty is that sex is not always available in administrative records. For example, Internal Revenue Service records do not have this variable except through the recoding of first names, which cannot be done with complete accuracy.

- *Date of Birth* – Day and month are generally well reported even by proxy respondents. Year can be used *with a tolerance* to good effect as a matching variable. Again, as with "sex", this item is not available on all administrative files.

Still other linkage variables could have been discussed, for example, race and telephone number. Race is a variable that is similar to sex except not nearly as well reported (unless it is recoded as black, nonblack, e.g., U.S. Bureau of the Census (1973)). Telephone numbers have problems similar to addresses and, while potentially of enormous value eventually, are not now widely available in administrative files.

3.2. Choice of a record-linkage system

One important aspect of the choice is whether to use model-based or ad hoc record-linkage techniques. As stated earlier, the authors favor model-based methods over ad hoc methods. The main drawback of ad hoc approaches is that they may not optimally treat the trade-offs that exist between cost and quality. For a given level of resources, therefore, they are unlikely to produce results as good as those that can be obtained by using model-based systems.

Model-based record linkage is likely to be technically more complex than an ad hoc approach and therefore requires a more extensive process of development. Fortunately, there now exist some proven general or multipurpose model-based systems (see Section 1) that are portable and reasonably well documented. This leads to the second important aspect of the choice: whether to use an existing record-linkage system or to develop a new customized system for the planned linkage. The first alternative offers the promise of avoiding many of the pitfalls typically met in designing record-linkage systems and therefore deserves full consideration. One of the major goals of the 1985 Workshop on Exact Matching Methodologies was to let potential users know about the characteristics and capabilities of available record-linkage systems.

A third important choice is whether the record-linkage system will be fully computerized or whether manual intervention will be permitted at some stages. Clearly, where large files are involved, many of the preprocessing steps and the basic matching steps like sorting, comparison of pairs, and calculation of weights must be fully automated. However, a review of current practice shows that manual intervention has a role in most systems. The system described by Howe and Lindsay (1981), for example, allows the visual inspection of records in connection with the establishment of threshold values (values of weights that serve as the boundaries between the decision categories $D(i)$ and the resolution of multiple matches). In describing the new record-linkage system being developed by the U.S. Census Bureau, Jaro (1985) refers to the resolution of possible links by a computer-assisted manual approach. He also proposes research on an iterative method of weight calculation that would require that comparison pairs that are not in full agreement be presented to an operator who would classify them as matches or nonmatches.

3.3. Specification of the desired outputs

To plan a record-linkage project it is, of course, necessary to specify clearly the desired final outputs. For many applications, it is desired that each comparison pair be assigned, explicitly or implicitly, to one of only two categories: positive link or positive nonlink. To reach this point without exceeding acceptable error levels, it may be necessary to inspect visually all or a sample of borderline comparison pairs (possible links) or even to collect additional data for some of them. Some method of resolving these cases must be specified.

The designer must also specify whether or not the existence of duplicates is considered possible in any of the files to be linked. If duplicates are possible, the system will have to include rules that permit multiple linkages as a final outcome. If they are not, a procedure will be needed to make final determinations when there are multiple linkages, i.e., the same record appears in more than one comparison pair that is classified initially as a positive link.

Some users may be satisfied with output files that simply provide data for the pairs classified as positive links and for the individual records not involved in any positive links. Others may wish to experiment with different threshold values, in which case it would be necessary to identify all comparison pairs with weights exceeding a specified threshold and to include the weight calculated for each such pair. Especially ambitious users might want to second guess some of the decisions made on blocking or calculating weights and try alternative methods. To make this possible, the outputs of the record-linkage process would have to include additional information. The 1973 Exact Match Study provided the documentation and tape files needed for this purpose (Aziz et al. (1978)).

In summary, the outputs should provide users with information that they can use to evaluate the record-linkage procedures and with the flexibility to use the linked data for a variety of purposes.

4. Preprocessing

Preprocessing consists of all procedures prior to the actual comparison of individual records: improving the quality of data in the input files, standardization of matching variables, and blocking of the input files. Each of these steps will be discussed.

4.1. Improving the quality of data in the input files

Matching errors can be minimized by doing as much as is feasible to ensure that the input files contain accurate information for the items that will or may be used as matching variables or for blocking. What can be done depends on how much control the record-linkage performer has over the development of the input files. Suppose, for example, that a statistical agency plans to do a survey and wants to enhance the results by linking the survey records with records from one or more administrative data systems. The statistical agency will have little, if any, control over the contents of the administrative files; at best it can devise procedures to detect and attempt to correct errors that exist in the files when received from their custodians. For the survey files, however, with sufficient advance planning, much can be done to ensure that data for the matching variables are complete, accurate, and in formats compatible with the corresponding items available from the administrative files (Scheuren (1983)).

These generalizations can be illustrated by considering the Social Security number (SSN), which is often used as a key matching variable in linking survey and administrative records in the United States. A first requirement is to do as much as possible to promote complete and accurate reporting of SSNs in the survey. Interviewer training can emphasize that the SSN is a legitimate and important survey item. Interviewers should be able to give accurate and reassuring answers to respondents who ask why SSNs are needed. Respondents can be asked to refer to records, especially if they are asked to report SSNs for other family members.

If the linkage of survey and administrative records is done in connection with a panel survey, one has the further recourse of collecting SSNs and other identifiers in the first round of interviews and using subsequent interviews to verify previously reported information and to seek additional information for persons for whom a positive link could not be established in the initial attempt at validation.

It has been demonstrated that procedures like those just described can be used to obtain validated SSNs for well over 90 percent of adult members of sample households in a well-managed survey (Kasprzyk (1983)).

What can be done in the less promising situation where one has no control over the input file and no access to the files of the Social Security Administration? This problem is equivalent to designing a computer edit of a survey data file at the stage where no further access to respondents or other external sources of data is feasible.

Continuing our use of the SSN for illustrative purposes, we find that both range checks and inter-item consistency checks can be used to detect incorrectly recorded SSNs. To date, roughly 30 percent of the one billion possible nine-digit SSNs have actually been issued. Using information made available by the Social Security Administration concerning the range of numbers already issued, some of the invalid SSNs can readily be detected. Consistency checks between date of birth and SSN are also possible, as explained by Jabine (1985). Invalid SSNs are sometimes the result of reporting errors, e.g., transposition of digits, and therefore need not necessarily be treated as missing values. Comparison pairs that agree on most digits of the SSN and show good agreement on other matching variables might be classified as positive or possible links. The appropriate treatment of cases in which two or more matching variables on a single record are known to be inconsistent is less obvious: the authors have not seen this question addressed in the literature on record linkage. Knowledge of the inconsistency in a record is clearly information that might be used in establishing the weights used for making linkage decisions.

4.2. Standardization of variables and formats

Standardization of matching variables can improve the chances of linking two records that match. In some cases, record linkage would be impossible without the initial reformatting of records: consider linking two files by name where one of the files has surnames first and the other has given names first. Differences in the presence or absence of titles (Mr., Mrs., Ms., Dr., etc.) can also cause difficulties. On the other hand, one must be aware that some kinds of standardization can result in distortion and loss of information and may increase the likelihood of designating some pairs of records as positive links when, in fact, they do not match.

Persons doing record linkages, of necessity, have given considerable attention to the standardization of names and addresses. Procedures vary depending on whether the units to be linked are persons or businesses.

For person names, a common procedure has been to encode surnames and to use the encoded values, usually along with other matching variables, for blocking. (The unencoded names may still be used in the matching step.) The two procedures most used are the Russell Soundex Code and the New York State Intelligence and Identification System (NYSIIS). The Soundex system, which is the older of the two, is a four-digit alpha-numeric code which uses the first letter of the surname, plus three digits based on the subsequent consonants in the surname (Jabine (1985)). The NYSIIS code is a variable length alphabetic code that divides the population of North American surnames into groups which vary less in size than those associated with the Soundex codes (Lynch and Arends (1977)). Both coding systems are designed so that surnames of similar sound have the same code and frequently-encountered errors of reporting do not cause changes in the code (Howe and Lindsay (1981)).

Other procedures sometimes used in formatting names include the removal of punctuation or blanks. For example, O'BRIEN becomes OBRIEN and VAN KAMP becomes VANKAMP. Dictionaries may be developed to relate commonly used nicknames to "official" names (BOB – ROBERT, BETSY – ELIZABETH) and to link common variations in spelling (SMITH – SMYTH – SMYTHE).

For business names, it is important to standardize abbreviations of commonly used terms, such as company, corporation, incorporated, limited, sales or distributor. This makes it possible, when records are compared, to distinguish these components of the name from those that are more likely to be unique to a business unit.

Standardization of addresses presents similar problems. There are several components which may appear in various sequences: street number, street name, apartment or other unit number, city or town, state and ZIP (postal) code. Some of these may be absent or replaced by other components such as rural delivery routes or post office boxes. To compare addresses on different records, these components must be identified and put in a standard format. The U.S. Census Bureau has developed a software package, called ZIPSTAN, for this purpose. ZIPSTAN was initially developed for use in geographic coding of addresses in the census of population. Examples of how ZIPSTAN formats addresses are given by Winkler (1985a).

One should be alert to the possibility of introducing errors when using manual procedures to standardize inputs. Marks (1985) pointed out that manual rearrangement and keying of unformatted names and addresses may introduce substantial error and that unaided computer formatting may introduce as much error as it removes. He describes an alternative technique in which clerks inserted a distinctive and computer-readable symbol in front of each of the name and address components to be used in matching, e.g., * before surname, # before house number, % before street name, etc.

There has been relatively little research on the effects of various standardization procedures on matching errors. Winkler (1985b) conducted experiments with a file of businesses containing known duplicates. He concluded that effective spelling standardization and accurate identification of corresponding subfields can help reduce matching errors.

4.3. Blocking and other procedures

Theoretical aspects of blocking have already been discussed in Section 2. As a practical matter, each of the input files must be sorted into blocks (sometimes referred to as "pockets") prior to the matching stage. Special provisions may be needed if multiple matches, each with a different blocking structure, are to be performed (see discussion in subsection 6.2, "Adjusting for false nonmatches"). It may also be desirable to assign sequence numbers to the records in each file so that comparison pairs can be readily identified.

5. Matching

The matching phase in a computerized record-linkage system generally consists of three stages: coding or classification of comparison pairs, preliminary decisions on linkages, and resolution of uncertainties. In probability-based systems the preliminary decisions are based on weights assigned to the observed outcomes of the comparisons; in ad hoc systems these decisions are usually based on predetermined rules relating comparison outcomes to the possible linkage decisions, $D(i)$. Most computerized systems require manual intervention at some stage, especially in resolving uncertainties that remain after the preliminary linkage decision step.

5.1. Coding comparison pairs

Within each block, every possible comparison pair (a, b) from the files A and B must be examined. In the examination of a pair, the normal procedure is to observe the extent of agreement or disagreement separately for each of the matching variables.

For any particular variable, the comparison outcomes can be coded or classified in a variety of ways. Suppose, for example, that one matching variable is race, and that every record in both files has one of two codes for race:

1 – black
2– nonblack

Comparison categories for this matching variable could be restricted to two:

1 – agree on race
2 – disagree on race

However, for reasons that will become evident later in this section, it may be desirable to use three categories:

1 – agree on race, race is black
2 – agree on race, race is nonblack
3 – disagree on race

For some matching variables, such as surname, much more complex coding systems may be used, and there are many possibilities. The comparison might be based on the actual surname, on a Soundex or NYSIIS encoded version, or on both. Some code schemes are based on a fixed number of letters, e.g., the first four letters (see Winkler (1985b), Howe and Lindsay (1981)). Still others make use of character-string comparison routines (string comparators) that take into account the likelihood of phonetic errors, transpositions of characters and random insertion, replacement and deletion of characters (Jaro (1985)).

Coding schemes for matching variables must also take into account the possibility that the variable will be missing for one or both members of the comparison pair.

Generally, if resources permit, all the variables judged suitable for linking should be used in the computer comparisons. When this is not possible, the variables can still be employed later in manually settling cases where the outcome might otherwise be indeterminate. However, manual intervention needs to be carefully limited and closely controlled. Manual matching is extremely costly and, while individual manual decisions can sometimes be better than those made by computer matching, usually humans lack consistency of judgment and can be distracted by extraneous information.

In most applications of the Fellegi-Sunter model the assumption is that agreement (or disagreement) on one linking variable is independent from that on any other, conditional only on whether or not the records brought together are for the same person. To aid in making this assumption plausible, special care needs to be taken in structuring agreement codes for variables like sex and first name that are inherently related (Fellegi (1985)). Kelley (1986) has done simulation studies to investigate the robustness of the U.S. Census Bureau's linkage system to violations of the independence assumption. For the particular population and linkage variables studied, he found that violations of the assumptions can have non-trivial effects on the levels of matching errors.

5.2. *Preliminary linkage decisions*

The heart of the linkage system is the procedure that assigns each comparison pair to one of two or more linkage categories, based on the codes assigned in the previous stage. In a probability-based system, the decision for each pair depends on the value of a linkage score or weight based on the comparison codes assigned for each matching variable. In the Fellegi-Sunter model, the weight is a function of the probability ratio:

$$\frac{\text{Prob (result of comparison, given match)}}{\text{Prob (result of comparison, given nonmatch)}}$$

The numerator represents the probability that the comparison of two records for the same person would produce the observed result. The denominator represents the probability that comparison of records for two different persons, selected at random, would produce the observed result. In general, the larger the ratio, the greater our confidence that the two records match, i.e., are for the same person.

To clarify this process, let us consider a simple example in which we are matching on both sex and race, where sex is always represented as either male or female and where race has been recoded black or nonblack. Further suppose that the proportion of males and females is each 50 percent and that blacks constitute 10 percent of the population and nonblacks 90 percent. Also suppose that the chance of a reporting or processing error for race is 1/100 and for sex 1/1000. Finally, we will assume that sex and race are independently distributed in the population and that reporting errors for matched pairs are independent.

With these stipulations and assumptions, we can calculate the probability or odds ratios shown in Table 1. Usually, given the independence assumption, the probability ratio is broken up into a series of ratios, one for each agreement or disagreement, and logs are taken (to the base 2 in this example). One is now working with simple sums, such that the larger (more positive) the total, the more likely that the pair is a match; conversely, the more negative the sum, the greater the likelihood that the two records are not for the same person.

*Table 1. Probability ratios for comparison outcomes based on race and sex**

Outcome of comparison	Probability ratio	Log of ratio (base 2)
Race and sex agree:		
Race is black	167.7369	7.3901
Race is nonblack	2.4589	1.2980
Race agrees, sex does not:		
Race is black	0.3358	- 1.5743
Race is nonblack	0.0049	- 7.6730
Sex agrees, race does not	0.2051	- 2.2856
Neither agree	0.0004	-11.2877

* See Computational Note (Appendix).

In this example it is only when both sex and race agree that the sum of the logs is positive. If the race is black, the log is between +7 and +8, moderately strong evidence in favor of a match. If the race is nonblack, however, the log is only slightly more than +1. As one would expect, the strongest evidence in favor of a nonmatch occurs when both race and sex disagree; for this outcome the log of the probability is about –11. This example illustrates nicely the fact that outcomes that are frequent in the population do not add very much to one's ability to decide if the pair should be treated as a link. However, if there are disagreements on such variables and reporting is reasonably accurate, then the variable may have a great deal of power in identifying comparison pairs that represent nonlinks.

In the Fellegi-Sunter model, the comparison pairs are ordered according to the values of their weights. Two cutoff points are established. The higher of these separates the positive links from the possible links and the lower one separates the possible links from the positive nonlinks.

The establishment of appropriate cutoff values is a critical part of any record linkage. The objective, under the Fellegi-Sunter model, is to place upper limits on the proportions of matched and unmatched pairs for which incorrect decisions are made. In choosing the target values, μ and λ, for these proportions, one should be aware that the number of unmatched pairs in the comparison space is usually much larger than the number of matched pairs. Therefore, it is usually desirable to make λ considerably smaller than μ; otherwise the false matches will tend to swamp the false nonmatches. Of course, the relative costs associated with the two kinds of errors may also influence the choice of μ and λ.

As we have seen, some applications of model-based record-linkage systems requires certain assumptions, such as independence of errors in the matching variables. Nevertheless, there is reason to believe that the Fellegi-Sunter procedure is fairly robust to departures from independence. Moderate errors in the estimation of weights can lead to different linkage decisions only for comparison outcomes whose weights are close to a cutoff point.

Furthermore, there is no theoretical obstacal to extending the underlying models to take into account known dependencies between linking variables (Kirkendall (1985)). There are also significant computational problems. Nevertheless, the approach is entirely workable, especially since the development of the Generalized Iterative Record Linkage System (GIRLS), which provides a state-of-the-art solution to the major computational problems (Smith and Silins (1981)). Other notable approaches in advanced linkage software include the work of Jaro and his collaborators (Jaro (1985)).

5.3. Resolving uncertainties

After the preliminary linkage decisions have been made, there are usually some uncertainties to be resolved. They consist primarily of pairs classified as possible links and of multiple links, i.e., groups of linked pairs that have one or more records in common.

In the Fellegi-Sunter procedure, possible links are the pairs that fall between the upper and lower cutoffs. If resources permit, these pairs may be reclassified as positive links or nonlinks either by collecting more data or by interactive manual review of the record content for these pairs.

If statistical estimates are to be made, and the resources needed to seek further information are not available, the potential links may be treated as nonlinks and a survey-type nonresponse adjustment may be made (Scheuren (1980)). It is possible, also, to consider keeping some of the potential links and then conducting the analysis, with an adjustment being made for mismatching (Scheuren and Oh (1975)).

Multiple links can occur in the Fellegi-Sunter formulation because the linkage decision is made independently for each pair. As a result, a record from either file may be included in more than one pair whose weight exceeds the cutoff for a positive link. In some applications, these many-to-one links might be appropriate, but usually a further step has to be taken to select the "best" one. This problem can occur with some frequency in administrative contexts. Manual review is usually the best basis for decisions, especially if further information is sought or is available to help make the selection. Some automated systems provide preliminary indications of the pairs judged to be the most likely candidates.

The National Death Index (NDI) operating system leaves it to users to resolve indeterminate cases. For each user record, they list as possible links all death records that qualify under one or more of 12 sets of matching criteria (e.g., agreement on SSN and first name, agreement on SSN and last name, agreement on month and day of birth and first and last names, etc.). NDI users with small files usually resolve multiple links by manual review. For large studies, some users have developed their own computer algorithms for this purpose (Patterson and Bilgrad (1985)). Users must also be prepared to determine final match status when only one possible link has been identified for a name submitted to the NDI. Doing this may often be more difficult than resolving multiple links.

Jaro (1985) offers a computerized transportation algorithm to solve multiple linkage problems. His approach is most effective when all the linking information has already been computerized and when there are contention problems in the linkages, that is, "*n*" records on one file are matching "*m*" records on another. The procedure is analogous to the "constrained matching" approach used in statistical matching, i.e., it picks a single best set of matches, rather than picking the best match for each record in one of the input files.

Even when many-to-one links are not appropriate in theory, it may be desirable to use the additional information they provide, especially if conditions do not permit a clear determination of which of the links represent true matches. Suppose, for example, that a record in File A, a sample file, is initially classified as a positive link with each of three records in File B, a 100 percent file. Three linked records could be established, each associating the File A record with one of the positive links from File B. The sample weight associated with the File A record would be divided among the three linked records: we might allocate one-third of it to each linked record or we might prefer to allocate it in proportion to the weights used in making the initial linkage decisions.

Difficulties with indeterminate cases can often be traced back to design flaws in the data linkage system. For example, not enough linking information may have been obtained on one or both files to assure uniqueness. The degree of redundancy in the identifiers may have been insufficient to compensate completely for the reporting errors. In an operational context, the linkage process may be so constrained by costs that, even if there are sufficient linkage items, they cannot be adequately exploited.

Some uncertainties may remain unresolved at this stage of the record linkage. Their resolution is discussed in the next section.

6. Analytical Issues

All too often researchers have embarked on large-scale record-linkage projects without full appreciation of the resources needed and the inevitability of matching errors. Typically, most of the resources allotted to such projects have been used up in the preprocessing and matching stages. The realization that matching errors remain at this point has sometimes led to disappointment with the results and even to their being thought completely unusable.

We believe there are solutions to this dilemma. Part of the answer, as discussed earlier under the heading "Specification of the desired outputs" (Section 3.3), is to give early attention to the purposes that the linked data files are to be used for and to structure the outputs accordingly, taking into account the need to deal with matching errors of both kinds. The presence of matching errors need not be viewed as an insurmountable obstacle. Survey researchers are accustomed to dealing with errors; they react by doing their best to create designs that minimize total error. They measure and document the remaining errors and consider their implications for the inferences that are drawn from the data. These same approaches can and should be used in performing record linkages, especially when their purpose is to enhance survey data with information from administrative sources.

Record-linkage systems, like survey-based or sample-based techniques, need to be "measurable" and to be made as robust as possible in the face of departures from the underlying assumptions. What can be done to achieve this is a sizable subject. We attempt here to sketch some of the issues and indicate general lines of attack.

6.1. Linkage documentation

Documentation should routinely be provided that tabulates the results of the match effort in terms that are relevant to the analysis. A distribution of the weights would be one example, perhaps shown for major subgroups. If a public-use file is being created, then the match weight might be placed in the file along with summary agreement codes, so that users can "second-guess" some of the decisions made. The inclusion of potential links, at least near the cutoff point, is another example of good practice. Most of these procedures were followed in the 1973 Exact Match Study (Aziz et al. (1978)), which linked U.S. data from the U.S. Current Population Survey with income tax and social security data.

6.2. *Adjusting for false nonmatches*

It is generally worthwhile to recompute the sample estimation weights for record pairs classified as positive links to adjust for false nonmatches (Scheuren (1980)). Conventional nonresponse procedures can be followed (Oh and Scheuren (1983)). Imputation strategies are also possible, but may be less desirable because they tend to disturb the estimated relationships between the two files being brought together (Oh and Scheuren (1980), Rodgers (1984)).

An important problem in this adjustment process, however conducted, is estimating whether a link should have occurred. Sometimes, by the nature of the problem, we know that all the records in one or both files should have been linked. In other cases (Rogot et al. (1983)), we are most interested in the linkage rate itself.

Elsewhere (Scheuren (1983)), we have advocated a capture-recapture approach to this estimation problem. Such an approach, in the presence of blocking, will allow us to improve the links obtained and will also make it possible to measure the extent to which our best efforts still lead to false nonmatches. Capture-recapture ideas are well described in the literature (e.g., Bishop et al. (1975), Marks et al. (1974)). Here we will only indicate the application.

If we employ two different blocking schemes and record the results of matching for every comparison pair determined to be a link under either scheme, we can display the results in the following contingency table:

Table 2. Links identified by the use of two different blocking schemes

Blocking scheme I	Blocking scheme II Link	Nonlink
Link	n_{11}	n_{12}
Nonlink	n_{21}	?

Some comparison pairs would not be examined under either blocking scheme and there could be some links included among them. By assuming that outcomes under the two blocking schemes are independent, we can estimate the number of linkages among these unexamined pairs, using the formula:

$$\hat{n}_{22} = \frac{n_{12}\ n_{21}}{n_{11}}$$

In practice, an analogous approach with three or more blocking schemes is recommended; otherwise the necessary assumptions may be unrealistically strong.

For best results, the blocks used in each scheme need to be as independent functionally and statistically as is possible, given the linkage information. Application of these ideas in nonstatistical uses of large administrative data files also seems worthy of study (Scheuren (1983)), although the expense of developing such an approach may be too great to incur unless there is a compelling administrative need.

Another approach to estimation of matching errors (of both kinds) is to test the record-linkage system with cases for which the true match status is known. Such testing would be relatively easy if, for example, we were working on a file of death records. To estimate false matches we could, as suggested by Smith (1985), match a set of records for persons known to be alive. To estimate false nonmatches we could use a set of records for persons known to have died. The latter method was used in the 1973 Exact Match Study to validate the manual search procedures used at the Social Security Administration (Kilss and Tyler (1974)). More recently, it has been used to evaluate alternative linkage rules for the National Death Index (Patterson and Bilgrad (1985)).

6.3. Adjusting for false matches

In most record-linkage studies, practitioners have operated in what they considered a conservative manner with regard to the links they should accept. Sometimes this may have meant heavy additional expense to obtain more information or the risk of seriously biasing results by leaving out a large number of the potential links. In any event, further research is needed on applying more complex analytic techniques that take explicit account of the false match rate, possibly by use of errors-in-variable approaches where the false match rate is estimated, e.g., as in Scheuren and Oh (1975). This would allow a correction factor to be derived. We must also attempt to find ways of estimating the false match rate that make weaker assumptions than those made in most Fellegi-Sunter applications.

In summary, the main issues in the analysis of linked data sets are that, at a minimum, we need to examine the sensitivity of the results to the assumptions made in the linkage process. Where possible, we need to quantify uncertainties in the results; specifically, indeterminacies in the linkages should translate into wider confidence intervals in the estimates. To achieve these goals we need to bring in techniques from other areas of statistics and apply them creatively to linked data sets. Examples here include information theory, errors-in-variable approaches and contingency table (capture-recapture) ideas.

7. Conclusions

The basic concepts of record linkage should be familiar to all of us. We apply them whenever we look for a number in a telephone directory. We have certain information: a name (we may be uncertain about spelling, middle initials, etc.), a place, and, possibly, a street and number. The scope of our search is limited by blocking techniques. We go to the directory for the appropriate geographic area and, in the latest U.S. directories, to the section for individuals, government agencies or businesses and professional organizations, as appropriate. In the section for individuals there is further blocking on last names.

If we cannot find a unique listing that is in full agreement with all of our matching variables, we look for listings that are in partial agreement and we place some of these in the possible link category. In this selection we make implicit judgments about how much weight to give to the observed comparison for each of the matching variables. These possible links are then resolved by telephoning, starting with the one deemed most promising, until a positive link is established.

Although the basic ideas are simple, there are many interesting and challenging technical problems that must be solved in undertaking large-scale record linkages. The record-linkage theory and models developed by Fellegi and Sunter and others offer valuable guides to solving these problems.

The 1985 Workshop on Exact Matching Methodologies provided a useful opportunity for people working in this area to share their techniques and experiences and to suggest areas in which additional research is needed. Following the workshop, the organizing committee developed a set of five recommendations for future research, development and evaluation activities (U.S. Internal Revenue Service (1985)). Because of their broad applicability, these recommendations are repeated here. While attendance at the workshop was limited to United States and Canadian specialists, the authors hope that others will be willing to join an informal network of persons interested in extending the application of model-based record-linkage procedures.

(1) Documentation should be improved and information on record linkage systems and techniques should be shared.

It is recommended that the Matching Group of the Administrative Records Subcommittee be reconstituted as a Technical Working Group on Record Linkage Systems and Techniques, continuing to function under the auspices of the Federal Committee on Statistical Methodology. The main goal of the Working Group would be to promote the effective use of record-linkage techniques for statistical purposes by encouraging the documentation of individual record-linkage systems and techniques and the sharing of relevant technical information. A primary activity would be sponsorship and organization of workshops and meetings of professional societies to discuss relevant new developments and research, and to disseminate information on existing systems and techniques. In addition, the reconstituted working group would contribute, in appropriate ways, to the implementation of recommendations 2 through 5 below.

(2) Changes in the external environment for record linkages should be monitored.

Statistical users of record-linkage techniques should track external developments that may influence their ability to perform record linkages. Such developments include changes in laws, regulations and policies affecting access to records and changes in the content of data files used in record linkages. Examples of the latter would include increased use of four-digit ZIP code add-ons ("ZIP + 4") and steps taken to promote the use of unique addresses in rural areas. In so far as possible, statistical users of record-linkage techniques, working through the reconstituted Working Group (see recommendation 1), should attempt to influence the course of these developments in ways that will facilitate statistical applications. For example, the Working Group might try to promote the development of standards for reporting names and addresses of both businesses and individuals.

(3) Comparative and evaluation studies of record-linkage systems should be undertaken.

Several agencies of the United States and Canadian governments have invested substantial resources in the development of automated record-linkage systems for use in a variety of statistical programs. For many new applications, use of an existing system is likely to be more cost-effective than the development of a new one. To aid potential users of record-linkage systems, we recommend that resources be sought for comparative evaluations of existing systems and some of their components, such as name and address standardizers and blocking rules. The evaluation design should recognize that record-linkage systems vary in their objectives, especially with respect to the kinds of units for which records are to be matched: persons or businesses. A much-needed first step is the development of a detailed evaluation plan that specifies the measures of quality and costs to be used in the evaluation and the nature of the files to be matched. Such evaluations may require data sets for which true match status is known. One possibility would be to create such data sets by simulation.

(4) Research and development aimed at the improvement of record-linkage systems and techniques should give priority to selected aspects.

Recognizing that resources for the development of improved record-linkage systems are limited, we recommend that priority be given to the following aspects: (1) systems for linking business records, (2) name and address standardizers, (3) string comparators, (4) the choice of blocking strategies, (5) the development of "learning" systems, and (6) the role of manual intervention.

(5) Errors associated with record linkages and their effects on analysis should be measured.

It is recommended that more research be

carried out on the error characteristics of record linkage systems and on the effects of errors on analyses performed with the linked data sets. To enhance the value of such research, consensus is desirable on standard measures of record linkage errors and on methods of measuring them. Promising error-measurement methods include multiple matching techniques and direct contacts with samples of linked pairs to determine their true match status.

Finally, as is well known to most readers, there are several important policy issues that have a bearing on the ability of agencies and researchers to perform record linkages for statistical purposes. Legal and ethical considerations must be weighed carefully by any organization that links records from different sources. Public perceptions of the appropriateness of various kinds of record linkages are of considerable importance (for an early discussion of these issues, see Steinberg and Pritzker (1967)).

One policy issue, in particular, arises directly from the improvements in record-linkage techniques that we have discussed here. Some statistical agencies release public-use microdata files (with explicit identifiers such as name and address removed) based on linked data sets, for use by researchers who wish to conduct their own analyses. The content of these public-use files is restricted in various respects to the point where it is believed that there is little likelihood that a user would be able, using externally available data, to identify one or more of the persons (or businesses) whose records were included in the file.

Unfortunately, the very techniques and tools that have made record linkages more feasible and efficient could, at least in theory, be used to attempt to identify persons included in public-use microdata files. Paass (1985), who has done research in this area, asserted that for many of the U.S. public files there are realistic scenarios in which disclosure of personal information seems to be possible. If he is correct, the implications for future release of such files are serious indeed. Statutory and administrative solutions, such as binding user agreements with penalties for violations and legal remedies for persons harmed by disclosure, may be the only way to ensure adequate protection to persons whose records are in these files.

Appendix

Computational Note

The probability ratios shown in Table 1 were calculated as follows:

Race and sex agree (Race is black)

$$[(.99)^2 + (.01)^2][(.999)^2 + (.001)^2]/[(.1)^2(.99)^2 + (.9)^2(.01)^2 + 2(.1)(.9)(.99)(.01)][.5] = 167.7369$$

Race and sex agree (Race is nonblack)

$$[(.99)^2 + (.01)^2][(.999)^2 + (.001)^2]/[(.9)^2(.99)^2 + (.1)^2(.01)^2 + 2(.9)(.1)(.99)(.01)[.5] = 2.4589$$

Race agrees, sex does not (Race is black)

$$[(.99)^2 + (.01)^2][2(.999)(.001)]/[(.1)^2(.99)^2 + (.9)^2(.01)^2 + 2(.1)(.9)(.99)(.01)][.5] = 0.3358$$

Race agrees, sex does not (Race is nonblack)

$$[(.99)^2 + (.01)^2][2(.999)(.001)]/[(.9)^2(.99)^2 + (.1)^2(.01)^2 + 2(.9)(.1)(.99)(.01)][.5] = 0.0049$$

Sex agrees, race does not

$$[2(.99)(.01)][(.999)^2 + (.001)^2]/[2(.1)^2(.99)(.01) + 2(.9)^2(.99)(.01) + 2(.9)(.1)(.99^2 + .01^2)][.5] = 0.2051$$

Neither agree

$$[2(.99)(.01)][2(.999)(.001)]/[2(.1)^2(.99)(.01) + 2(.9)^2(.99)(.01) + 2(.9)(.1)(.99^2 + .01^2)][.5] = 0.0004$$

8. References

Alvey, W. and Aziz, F. (1979): Mortality Reporting in SSA Linked Data: Preliminary Results. Social Security Bulletin, 42 (11), pp. 15–19.

Arellano, M. (1985): An Implementation of a Two-Population Fellegi-Sunter Probability Linkage Model. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 255–258.

Aziz, F., Kilss, B., and Scheuren, F. (1978): 1973 Current Population Survey – Administrative Record Exact Match File Codebook, Part I, Code Counts and Item Definitions. Studies from Interagency Data Linkages, U.S. Social Security Administration, No. 8.

Bishop, Y., Fienberg, S., and Holland, P. (1975): Discrete Multivariate Analysis. Cambridge: MIT Press.

Chandrasekaran, C. and Deming, W. (1949): On a Method of Estimating Birth and Death Rates and the Extent of Registration. Journal of the American Statistical Association, 44, pp. 101–115.

Childers, D. and Hogan, H. (1984): Matching IRS Records to Census Records: Some Problems and Results. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 301–306.

Cobleigh, C. and Alvey, W. (1975): Validating the Social Security Number. Studies from Interagency Data Linkages, U.S. Social Security Administration, No. 4, pp. 89–123.

Coulter, R. (1985): An Application of a Theory for Record Linkage. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 89–96.

Fellegi, I. (1985): Tutorial on the Fellegi-Sunter Model for Record Linkage. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 127–138.

Fellegi, I. and Sunter, A. (1969): A Theory for Record Linkage. Journal of the American Statistical Association, 64, pp. 1183–1210.

Howe, G. and Lindsay, J. (1981): A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies. Computers and Biomedical Research, 14, pp. 327–340.

Jabine, T. (1985): Properties of the Social Security Number Relevant to Its Use in Record Linkages. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 213–225.

Jaro, M. (1972): Unimatch – A Computer System for Generalized Record Linkage Under Conditions of Uncertainty. AFIPS, Conference Proceedings.

Jaro, M. (1985): Current Record Linkage Research. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 317–320.

Kasprzyk, D. (1983): Social Security Number Reporting, the Use of Administative Records and the Multiple Frame Design in the Income Survey Development Program. Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program, Social Science Research Council, pp. 123–144.

Kelley, R. (1984): Blocking Considerations for Record Linkage Under Conditions of Uncertainty. American Statistical Association, Proceedings of the Social Statistics Section, pp. 602–605.

Kelley, R. (1985): Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 199–203.

Kelley, R. (1986): Robustness of the Census Bureau's Record Linkage System. Presented at the August 1986 meeting of the American Statistical Association.

Kilss, B. and Scheuren, F. (1978): The 1973 CPS-IRS-SSA Exact Match Study. Social Security Bulletin 41(10), pp. 14–22.

Kilss, B. and Tyler, B. (1974): Searching for Missing Social Security Numbers. American Statistical Association, Proceedings of the Social Statistics Section, pp. 137–144.

Kirkendall, N. (1985): Weights in Computer Matching: Applications and an Information Theoretic Point of View. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 189–198.

Kullback, S. (1968): Information Theory and Statistics. New York: Dover.

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

Marks, E. (1985): Discussion (of paper by W. Winkler). Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 205–206.

Marks, E., Seltzer, W., and Krotki, K. (1974): Population Growth Estimation: A Handbook of Vital Statistics Measurement. New York: The Population Council.

Newcombe, H., Kennedy, J., Axford, S., and James, A. (1959): Automatic Linkage of Vital Records. Science, 130 (3381), pp. 954–959.

Oh, H. and Scheuren, F. (1980): Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 416–420.

Oh, H. and Scheuren, F. (1983): Weighting Adjustments for Unit Nonresponse. Incomplete Data in Sample Surveys (Vol. 2), Panel on Incomplete Data, U.S. National Academy of Sciences, pp. 143–184.

Paass, G. (1985): Disclosure Risk and Disclosure Avoidance for Microdata. Presented at the May 1985 meetings of the International Association for Social Service Information and Technology (IASSIST).

Patterson, J. and Bilgrad, R. (1985): The National Death Index Experience: 1982–1985. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 245–254.

Rodgers, W. (1984): An Evaluation of Statistical Matching. Journal of Business and Economic Statistics, 2, pp. 91–102.

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

Scheuren, F. (1980): Methods of Estimation for the 1973 Exact Match Study. Studies from Interagency Data Linkages, U.S. Social Security Administration, No. 10, pp. 1–123.

Scheuren, F. (1983): Design and Estimation for Large Federal Surveys Using Administrative Records. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 377–381.

Scheuren, F. and Herriot, R. (1975): The Role of the Social Security Number in Matching Administrative and Survey Records – General Introduction and Background. Studies from Interagency Data Linkages, U.S. Social Security Administration, No. 4, pp. 1–7.

Scheuren, F. and Oh, H. (1975): Fiddling Around with Nonmatches and Mismatches. American Statistical Association, Proceedings of the Social Statistics Section, pp. 627–633.

Smith, M. (1985): Record-Keeping and Data Preparation Practices to Facilitate Record Linkages. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 321–326.

Smith, M. and Silins, J. (1981): Generalized Iterative Record Linkage System. American Statistical Association, Proceedings of the Social Statistics Section, pp. 128–137.

Steinberg, J. and Pritzker, L. (1967): Some Experiences With and Reflections on Data Linkage in the United States. Bulletin of the International Statistical Institute 42(2), pp. 786–808.

Tepping, B. (1968): A Model for Optimum Linkage of Records. Journal of the American Statistical Association, 63, pp. 1321–1332.

U.S. Bureau of the Census (1973): The Medicare Record Check: An Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1979 Census, Evaluation and Research Program, Series PC (E), No. 7.

U.S. Department of Health, Education and Welfare (1973): Records, Computers, and the Rights of Citizens. Report of the Secretary's Advisory Committee on Automated Personal Data Systems.

U.S. Internal Revenue Service (1985): Record Linkage Techniques-1985. Proceedings of the Workshop on Exact Matching Methodologies, May 9–10, 1985.

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

Winkler, W. (1985a): Preprocessing of Lists and String Comparison. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 181–188.

Winkler, W. (1985b): Exact Matching Lists of Businesses: Blocking, Subfield Identification and Information Theory. Record Linkage Techniques-1985, U.S. Internal Revenue Service, pp. 227–241.

Received April 1986
Revised September 1986