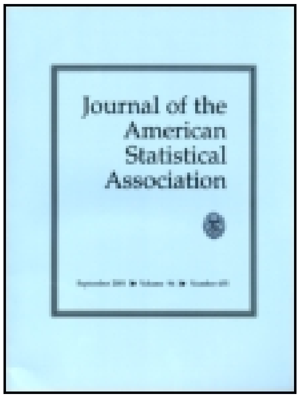
This article was downloaded by: [Stephen E. Fienberg]
On: 14 July 2013, At: 13:43
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems

Mauricio Sadinle [a] & Stephen E. Fienberg [b]
[a] Department of Statistics , Carnegie Mellon University , Pittsburgh , PA , 15213
[b] Department of Statistics, the Machine Learning Department, and the Heinz College , Carnegie Mellon University , Pittsburgh , PA , 15213

Accepted author version posted online: 02 Jan 2013.Published online: 01 Jul 2013.

To cite this article: Mauricio Sadinle & Stephen E. Fienberg (2013) A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems, Journal of the American Statistical Association, 108:502, 385-397, DOI: 10.1080/01621459.2012.757231

To link to this article: http://dx.doi.org/10.1080/01621459.2012.757231

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

A Generalized Fellegi–Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems

Mauricio SADINLE and Stephen E. FIENBERG

We present a probabilistic method for linking multiple datafiles. This task is not trivial in the absence of unique identifiers for the individuals recorded. This is a common scenario when linking census data to coverage measurement surveys for census coverage evaluation, and in general when multiple record systems need to be integrated for posterior analysis. Our method generalizes the Fellegi–Sunter theory for linking records from two datafiles and its modern implementations. The goal of multiple record linkage is to classify the record K-tuples coming from K datafiles according to the different matching patterns. Our method incorporates the transitivity of agreement in the computation of the data used to model matching probabilities. We use a mixture model to fit matching probabilities via maximum likelihood using the Expectation–Maximization algorithm. We present a method to decide the record K-tuples membership to the subsets of matching patterns and we prove its optimality. We apply our method to the integration of the three Colombian homicide record systems and perform a simulation study to explore the performance of the method under measurement error and different scenarios. The proposed method works well and opens new directions for future research.

KEY WORDS: Bell number; Census undercount; Data linkage; Data matching; EM algorithm; Mixture model; Multiple systems estimation; Partially ordered set.

1. INTRODUCTION

Record linkage is a widely used technique for identifying records that refer to the same individual across different datafiles. This task is not trivial when unique identifiers are not available, and many authors have proposed probabilistic methods to deal with this problem building upon the seminal work of Newcombe et al. (1959) and Fellegi and Sunter (1969). Applications of record linkage include merging post-enumeration surveys and census data for census coverage evaluation (e.g., Winkler 1988; Jaro 1989; Winkler and Thibaudeau 1991), linking health care databases for epidemiological studies (e.g., Bell, Keesey, and Richards 1994; Méray et al. 2007), and adaptive name matching in information integration (Bilenko et al. 2003) among others.

1.1 Linking Multiple Datafiles

To perform record linkage involving more than two datafiles, some authors have used record linkages for each pair of datafiles or other ad hoc procedures (see, e.g., Darroch et al. 1993; Zaslavsky and Wolfgang 1993; Asher and Fienberg 2001; Asher et al. 2003; Méray et al. 2007). Separate pairwise matchings of datafiles do not guarantee the transitivity of the linkage decisions and thus require resolving discrepancies (Fienberg and Manrique-Vallier 2009). For example, let us suppose we link the record of the individual a in a first datafile and the record of an individual b in a second datafile from a bipartite record linkage (classical record linkage of two datafiles). Then, from a second bipartite record linkage, we link the record of b to the record of an individual c in a third datafile. Based on these two linkages we might conclude that a, b, and c are the same individual. Unfortunately, had we also linked the first and third files, a and c may not match. If a, b, and c truly correspond to the same individual, the nonmatch could occur due to measurement error or incomplete record information. On the other hand, if the records of a, b, and c do not refer to the same individual, we have four possibilities: a and b refer to the same individual but c refers to another one, a and c refer to the same individual but b refers to another one, b and c refer to the same individual but a refers to another one, or all a, b, and c refer to different individuals. By using bipartite record linkage for each pair of files, we cannot resolve the matching pattern for these three records. While there are various ad hoc approaches to resolve the results of multiple bipartite matchings, no formal methodology has appeared in the statistical literature (see, e.g., the recent surveys of Herzog, Scheuren, and Winkler 2007, 2010).

1.2 Census and Record Systems Coverage Evaluation

Implementation of accurate methods for census coverage evaluation and possibly census adjustment requires the integration of multiple datafiles. The usual methodology of census coverage evaluation matches a coverage measurement survey to the census data to estimate population sizes using dual-system estimation (Hogan 1992, 1993). This procedure is subject to "correlation bias" that results when responses to the census and

---

Mauricio Sadinle is Ph.D. student, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *msadinle@stat.cmu.edu*). Stephen E. Fienberg is Maurice Falk University Professor of Statistics and Social Science in the Department of Statistics, the Machine Learning Department, and the Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *fienberg@stat.cmu.edu*). This research was partially supported by NSF grants BCS-0941518 and SES-1130706 to Carnegie Mellon University, and by the Singapore National Research Foundation under its International Research Centre at Singapore Funding Initiative and administered by the IDM Programme Office. The authors thank Rob Hall, Kristian Lum, Michael Larsen, the Associate Editor, and two referees for helpful comments and suggestions on earlier versions of this article, and Jorge A. Restrepo for providing the Colombian homicide data. An early version of this article was written by the first author when he was affiliated to the Conflict Analysis Resource Center (CERAC) and the National University of Colombia at Bogotá.

Journal of the American Statistical Association
June 2013, Vol. 108, No. 502, Applications and Case Studies
DOI: 10.1080/01621459.2012.757231

survey are dependent or the joint inclusion probabilities are heterogeneous (Darroch et al. 1993; Zaslavsky and Wolfgang 1993; Anderson and Fienberg 1999). The incorporation of additional surveys or administrative data into the coverage evaluation process allows for checking on assumptions regarding independence of lists and homogeneity, and for modeling departures from them. This in turn requires attention to the problem of multiple record linkage.

Likewise, underregistration is the norm rather than the exception in record systems of human rights violations and violent events in general, especially in countries with high levels of violence. Discrepancies appear whenever there are different record systems capturing information about the same event of interest. The diversity of sources provides a useful input for coverage assessment of the different record systems (e.g., Ball 2000; Gohdes 2010; Lum et al. 2010). A clear example of this scenario occurs in Colombia, where there exist three homicide record systems that usually differ in the number of recorded casualties. Those record systems are maintained by the Colombian Census Bureau (Departamento Administrativo Nacional de Estadistica – DANE, in Spanish), the Colombian National Police (Policia Nacional de Colombia), and the Colombian Forensics Institute (Instituto Nacional de Medicina Legal y Ciencias Forenses). The discrepancies in the numbers recorded by these record systems are the result of conceptual and methodological differences among these institutions, as well as problems of geographical coverage (Restrepo and Aguirre 2007). Whereas the data from the National Police and Forensics Institute simply record the information obtained from their daily activities, the objective of the Colombian Census Bureau is to determine the true number of deaths occurring in Colombia and its geographical subdivisions (Departamento Administrativo Nacional de Estadisticas, DANE 2009). Thus, the coverage evaluation of the Colombian Census Bureau record system is important, and its linkage with the other two sources can lead to improved estimates of the number of homicides.

1.3 Overview of the Article

We propose a method for the linkage of multiple datafiles, generalizing the theory of Fellegi and Sunter (1969) and the implementations presented by Winkler (1988) and Jaro (1989), which still represent the mainstream approach for unsupervised record linkage (see Copas and Hilton 1990 for a supervised approach). Our method incorporates the transitivity of agreement in the computation of the data used to model matching probabilities. In Section 2, we generalize the set of record pairs presented by Fellegi and Sunter (1969) to a K-ary product of the K datafiles to be linked, and we present this K-ary product as the union of all the possible subsets that contain the possible patterns of agreement of the record K-tuples. In Section 3, we propose a method to compute comparison data from record K-tuples, incorporating transitivity, and we present a way to schematize this kind of data through simple graphs. To fit matching probabilities, in Section 4 we generalize the mixture model used by Winkler (1988) and Jaro (1989), and in Section 5 we present details of the fitting of this model using the Expectation–Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977). In Section 6, we present an optimal method to decide the record K-tuples membership to the subsets defined in Section 2. Section 7 contains an application of the proposed methods to the integration of the three Colombian homicide record systems, and Section 8 describes simulation studies where we explore the performance of the method under different scenarios.

2. COVERED SUBPOPULATIONS AND RECORD K-TUPLES

We follow the exposition of Fellegi and Sunter (1969) and suppose some population is recorded by K datafiles. Let $A_1, A_2, \ldots, A_K$ denote the K overlapping subpopulations recorded in those K datafiles. Now, suppose that for each datafile there exists one different record generating process α_k, which produces a set of records denoted by

$$\alpha_k(A_k) = \{\alpha_k(a_k); a_k \in A_k\}, k = 1, \ldots, K,$$

where the member $\alpha_k(a_k)$ represents a vector of information of the member $a_k \in A_k$. This information could be subject to measurement error or incomplete. Let us define the K-ary Cartesian product

$$\bigotimes_{k=1}^{K} \alpha_k(A_k) = \{(\alpha_1(a_1), \alpha_2(a_2), \ldots, \alpha_K(a_K)); a_k \in A_k, \quad k = 1, \ldots, K\},$$

which is composed by all the possible record K-tuples in which the kth entry corresponds to the information recorded for some a_k in the subpopulation k. Now we describe the possible matching patterns of the record K-tuples in terms of the members of the subpopulations A_k. First, it is possible that a record K-tuple includes information on K different individuals, that is, for some $(\alpha_1(a_1), \alpha_2(a_2), \ldots, \alpha_K(a_K))$, $a_k \neq a_{k'}$, for all $k \neq k'$. At the other extreme, if an individual appears in all K datafiles, then in the record K-tuple $(\alpha_1(a_1), \alpha_2(a_2), \ldots, \alpha_K(a_K))$ actually $a_1 = a_2 = \ldots = a_K$. In general, we can classify the entries of each record K-tuple into subsets that record information on the same individual.

To establish this idea formally, let $\mathbb{P}_K$ denote the set of partitions of the set $\mathbb{N}_K = \{1, 2, \ldots, K\}$. If we associate each number in $\mathbb{N}_K$ with an entry in a record K-tuple, then the matching pattern of each record K-tuple corresponds to a partition of $\mathbb{N}_K$, where the elements of the partition group the entries of the K-tuple that represent the same individual. Now, let S_p denote the set of record K-tuples corresponding to the matching pattern $p \in \mathbb{P}_K$. It is clear that

$$\bigotimes_{k=1}^{K} \alpha_k(A_k) = \bigcup_{p \in \mathbb{P}_K} S_p \tag{1}$$

since each record K-tuple has a unique matching pattern. The number of ways we can partition a set of K elements into nonempty subsets is called the Kth Bell number, denoted B_K, which can be found using the recurrence relation $B_K = \sum_{k=0}^{K-1} B_k \binom{K-1}{k}$, with $B_0 = 1$ by convention (see Rota 1964 for further details). Thus, there are B_K subsets S_p of record K-tuples.

Let n denote the cardinality of the set in Equation (1). Also, for $j = 1, \ldots, n$, let $r_j = (\alpha_1(a_1), \ldots, \alpha_K(a_K))$ for some $a_k \in A_k$, $k = 1, \ldots, K$, be the jth record K-tuple of the K-ary product

Table 1. Each matching pattern of a record triplet can be associated with a partition of the set $\{1, 2, 3\}$

Notation	$\mathbb{P}_3$	$(\alpha_1(a_1), \alpha_2(a_2), \alpha_3(a_3))$
1/2/3	$\{\{1\}, \{2\}, \{3\}\}$	$a_1 \neq a_2 \neq a_3 \neq a_1$
12/3	$\{\{1, 2\}, \{3\}\}$	$a_1 = a_2; a_3 \neq a_1, a_2$
13/2	$\{\{1, 3\}, \{2\}\}$	$a_1 = a_3; a_2 \neq a_1, a_3$
1/23	$\{\{1\}, \{2, 3\}\}$	$a_2 = a_3; a_1 \neq a_2, a_3$
123	$\{\{1, 2, 3\}\}$	$a_1 = a_2 = a_3$

in Equation (1). When the datafiles do not contain common identifiers, we cannot identify the subset S_p to which the record K-tuple r_j belongs. If the datafiles record the same F fields of information, however, we can obtain a comparison vector γ^j for each K-tuple r_j. We can use this information to estimate the probability that each record K-tuple belongs to each subset S_p, given the comparison vector γ^j. Multiple record linkage's goal is to classify all the record K-tuples in the appropriate subsets S_p.

Example 1. If we have $K = 3$ datafiles, for each triplet of records we have the matching patterns in Table 1, which can be represented using undirected graphs as in Figure 1. In this case, we also have $B_3 = 5$ and the Cartesian product of the three datafiles can be written as

$$\bigotimes_{k=1}^{3} \alpha_k(A_k) = S_{1/2/3} \cup S_{12/3} \cup S_{13/2} \cup S_{1/23} \cup S_{123}. \quad (2)$$

2.1 Blocking

Note that the dimension of the K-ary product grows exponentially as a function of K. Thus, considering the complete set of record K-tuples is highly inefficient in most applications. A common way to deal with this problem in bipartite record linkage is to partition each datafile into a common set of blocks, thereby eliminating the need to match records in different blocks. The idea is that reliable categorical fields such as zip code or gender may be used to quickly label some of the nonlinks. For example, if we are matching datafiles with geographic information, we could assign those records that differ in zip code (or a similar field) as nonlinks. See Herzog, Scheuren, and Winkler (2007, 2010) and Christen (2012) for a discussion of blocking.

In multiple record linkage, we can apply the same idea to assign nonlinks between pairs of records within every record K-tuple. If a certain blocking variable assigns a nonlink between records k and k' in the record K-tuple r_j, this implies that r_j cannot be assigned to subsets S_p where the pattern of agreement p involves a link between files k and k'. Consequently, the record linkage process has to decide among the remaining possibilities. If a nonlink is assigned to every pair of records within a record K-tuple, then this K-tuple can be assigned directly to the subset $S_{1/2/\ldots/K}$ (see notation in Table 1). In practice this last step tremendously reduces the number of K-tuples to be classified.

Using the natural partial order in $\mathbb{P}_K$, we provide a way to determine the subsets to which a record K-tuple can be assigned after blocking. We say that $p' \preccurlyeq p$ if p' is a partition finer than or equal to p. Note that the blocking process provides a maximal pattern of agreement p_b for each record K-tuple r_j. Thus, the subsets to which r_j can be potentially assigned are those S_p such that $p \preccurlyeq p_b$.

Example 2. In Figure 2, we present the Cartesian product of two pairs of files after blocking. We illustrate using homicide data from the Armenia, Montenegro, and Quimbaya towns in the Colombian province of Quindio. In this example, only the gray elements of the Cartesian product become part of the record linkage process, whereas the white elements become a priori nonmatches. The left-hand side of Figure 2 represents the Cartesian product of two Census and Police data subsets after blocking by town. The right-hand side represents the Cartesian product of the same Census data subset and a Forensics data subset after blocking by gender. Note that in this example we assign the pair $(\alpha_1(a), \alpha_2(b))$ as a nonlink since these two records refer to homicides in different towns. We also assign the pair $(\alpha_1(a), \alpha_3(c))$ as a nonlink since these two records refer to different genders. Assuming that there are no nonlink blocking assignments for $(\alpha_2(b), \alpha_3(c))$, the multiple record linkage decision process has to classify the triplet $(\alpha_1(a), \alpha_2(b), \alpha_3(c))$ as either belonging to $S_{1/2/3}$ or $S_{1/23}$. On the other hand, the two blocking processes illustrated in Figure 2 have no direct implications on the possible resolution of $(\alpha_1(d), \alpha_2(b), \alpha_3(c))$.

3. COMPARISON DATA

To obtain appropriate data to model the probability that a certain record K-tuple belongs to some subset S_p, let us determine the matching pattern for each common field of recorded information. If for a certain record K-tuple we search for agreement among the information recorded for a certain field, we can associate each entry of the record K-tuple with a number in $\{1, 2, \ldots, K\}$ and a certain partition of this set would describe the matching pattern of the record K-tuple for the field in consideration, grouping in the same element of the partition all the K-tuple entries that agree in the field being compared (similar to Section 2). An alternative way to explain this idea is as follows. For some record K-tuple, let us compare the information of the records from the datafiles k, k', and k'' for a certain common field. Due to transitivity of agreement, if records k and k' agree and k' and k'' agree, then k and k'' agree necessarily.

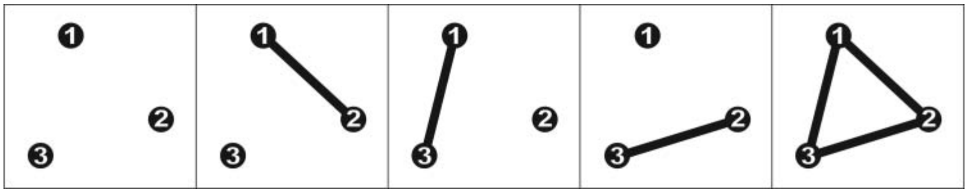


Figure 1. Undirected graphs giving $B_3 = 5$ possible patterns of agreement using three datafiles. The vertices appear connected if the value that each one represents agree, otherwise, the vertices appear unconnected.

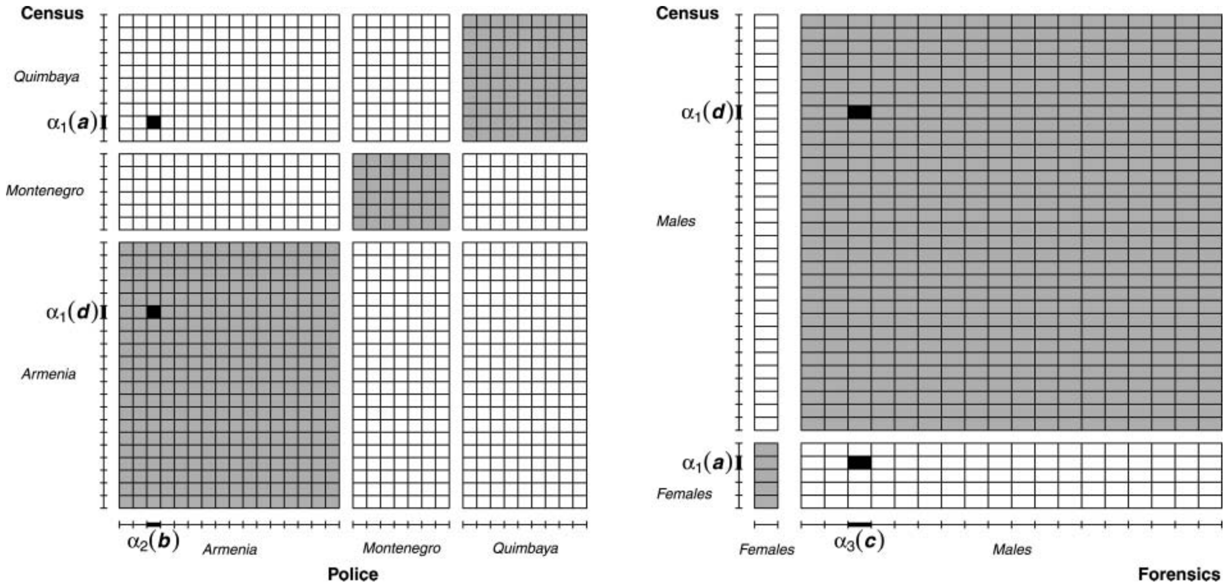


Figure 2. Cartesian products of Census and Police homicide data after blocking by town (left), and Census and Forensics homicide data after blocking by gender (right) for three towns in Colombia. Only elements in gray blocks are potentially linked. Black elements are discussed in the example of Section 2.1.

Thus, since agreement is an equivalence relation, each matching pattern for each field for each record K-tuple is a partition of K points, because for any equivalence relation on a set, the set of its equivalence classes (sets of records agreeing) is a partition of the set.

Now, let $\gamma_p^{j_f} = 1$ if the record K-tuple r_j has the matching pattern p in the field f. Then, for each field $f = 1, \ldots, F$, of each record K-tuple r_j, we obtain a vector $\gamma^{j_f} = (\gamma_{1/2/\ldots/K}^{j_f}, \ldots, \gamma_p^{j_f}, \ldots, \gamma_{12\ldots K}^{j_f})$, where only one entry is equal to one and the rest are equal to zero. Note the length of the vector γ^{j_f} is B_K, since this is the number of patterns of agreement for each field. Finally, the comparison data for r_j contains the comparison vectors for all the F fields, and can be written as $\gamma^j = (\gamma^{j_1}, \ldots, \gamma^{j_f}, \ldots, \gamma^{j_F})$, which takes values over $(B_K)^F$ possible matching patterns.

Similarly as in Section 2, we can represent the patterns of agreement presented in this section by unions of complete undirected graphs (see Rosen 2006, p. 448) as in Figure 1. In those graphs, each vertex represents the value of certain field in certain record that belongs to certain datafile $k = 1, \ldots, K$. The vertices k' and k appear connected if the values that they represent agree, otherwise, the vertices appear disconnected.

Example 3. Let us expose how the comparison data work when we need to link three datafiles. In this case, we can represent the patterns of agreement as five unions of complete undirected graphs, as presented in Figure 1. For $K = 3$, $\gamma^{j_f} = (\gamma_{1/2/3}^{j_f}, \gamma_{12/3}^{j_f}, \gamma_{13/2}^{j_f}, \gamma_{1/23}^{j_f}, \gamma_{123}^{j_f})$ represents the comparison data for the field f (say age, ethnicity, etc.) of the record triplet r_j, and the length of the full comparison data for each record triplet is $5F$, if the datafiles have F common fields.

4. MODEL FOR MATCHING PROBABILITIES

The probabilities $P(S_p|\gamma^j) \doteq P(r_j \in S_p|\gamma^j)$, $p \in \mathbb{P}_K$, can be found using $P(\gamma^j|S_p) \doteq P(\gamma^j|r_j \in S_p)$ and $P(S_p) \doteq P(r_j \in S_p)$, as $P(S_p|\gamma^j) = P(\gamma^j|S_p)P(S_p)/P(\gamma^j)$, where

$$P(\gamma^j) = \sum_{p \in \mathbb{P}_K} P(\gamma^j|S_p)P(S_p).$$

Let $g^j = (g_{1/2/\ldots/K}^j, \ldots, g_{12\ldots K}^j)$ be the vector that indicates the subset S_p that contains the record K-tuple r_j, such that $g_p^j = 1$ if $r_j \in S_p$ and $g_p^j = 0$ otherwise. Thus, it is clear that $\sum_{\mathbb{P}_K} g_p^j = 1$. Now, let $x^j = (g^j, \gamma^j)$ be the (partially observed) complete data vector for r_j. Note that after blocking, some entries of g^j are fixed as zeroes for some record K-tuples.

Winkler (1988), Jaro (1989), and Larsen and Rubin (2001) proposed to model the corresponding complete data x^j for bipartite record linkage, where g^j is taken as a latent variable. For multiple record linkage, the model for x^j is stated as

$$P(x^j|\Phi) = \prod_{p \in \mathbb{P}_K} [P(\gamma^j|S_p)P(S_p)]^{g_p^j}.$$

Under the conditional independence assumption of the comparison data fields, we obtain

$$P(\gamma^j|S_p) = \prod_{f=1}^{F} P(\gamma^{j_f}|S_p). \tag{3}$$

Each γ^{j_f} represents the matching pattern of r_j in the field f, which corresponds to categorical information that can be modeled by using a categorical distribution (or multinomial with just one trial) as

$$P(\gamma^{j_f}|S_p) = \prod_{p' \in \mathbb{P}_K} \left(\pi_{p'|p}^f\right)^{\gamma_{p'}^{j_f}}, \tag{4}$$

where $\pi_{p'|p}^f \doteq P(\gamma_{p'}^{j_f} = 1|S_p)$, and p' is just another indicator of the patterns of agreement in $\mathbb{P}_K$. Defining $\mathscr{s}_p \doteq P(S_p)$, under independence of the complete data, the complete log-likelihood for the sample $\mathbf{x} = \{x^j; j = 1, \dots, n\}$ is obtained as

$$L = \sum_{j=1}^{n} \sum_{p\in\mathbb{P}_K} g_p^j \left[\log \mathscr{s}_p + \sum_{f=1}^{F} \sum_{p'\in\mathbb{P}_K} \gamma_{p'}^{j_f} \log \pi^f_{p'|p} \right].$$

The set of parameters in the log-likelihood above is $\Phi = (\mathbf{s}, \Pi)$, where $\mathbf{s}$ is a vector of length B_K given by $\mathbf{s} = (\mathscr{s}_{1/2/\dots/K}, \dots, \mathscr{s}_{12\dots K})$ and Π can be arranged in a set of F matrices of size $B_K \times B_K$, each one given by

$$\Pi^f = \begin{pmatrix} \pi^f_{1/2/\dots/K|1/2/\dots/K} & \cdots & \pi^f_{1/2/\dots/K|p} & \cdots & \pi^f_{1/2/\dots/K|12\dots K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \pi^f_{p'|1/2/\dots/K} & \cdots & \pi^f_{p'|p} & \cdots & \pi^f_{p'|12\dots K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \pi^f_{12\dots K|1/2/\dots/K} & \cdots & \pi^f_{12\dots K|p} & \cdots & \pi^f_{12\dots K|12\dots K} \end{pmatrix}$$

for $f = 1, \dots, F$. Hence, the length of Φ is $B_K(B_K F + 1)$. To estimate these probabilities, since the g^j vectors are only partially observed, the estimation is made via maximum likelihood using the EM algorithm (Dempster, Laird, and Rubin 1977). The model presented in this section generalizes the one used by Winkler (1988) and Jaro (1989), and uses the strong assumption that the comparison data fields are conditionally independent given the K-tuples' membership to the subsets S_p. In Section 7, we show that this baseline model produces good results for the Colombian homicide data, but the modeling of the fields' dependencies may be a key factor in obtaining good linkage results in other contexts (see Larsen and Rubin 2001). This is part of our ongoing work.

Example 4. For the particular case where $K = 3$, the length of Φ is $5 + 25F$, which is given by $\mathbf{s} = (\mathscr{s}_{1/2/3}, \mathscr{s}_{12/3}, \mathscr{s}_{13/2}, \mathscr{s}_{1/23}, \mathscr{s}_{123})$ and Π, which is composed by F matrices of size 5×5, as

$$\Pi^f = \begin{pmatrix} \pi^f_{1/2/3|1/2/3} & \pi^f_{1/2/3|12/3} & \pi^f_{1/2/3|13/2} & \pi^f_{1/2/3|1/23} & \pi^f_{1/2/3|123} \\ \pi^f_{12/3|1/2/3} & \pi^f_{12/3|12/3} & \pi^f_{12/3|13/2} & \pi^f_{12/3|1/23} & \pi^f_{12/3|123} \\ \pi^f_{13/2|1/2/3} & \pi^f_{13/2|12/3} & \pi^f_{13/2|13/2} & \pi^f_{13/2|1/23} & \pi^f_{13/2|123} \\ \pi^f_{1/23|1/2/3} & \pi^f_{1/23|12/3} & \pi^f_{1/23|13/2} & \pi^f_{1/23|1/23} & \pi^f_{1/23|123} \\ \pi^f_{123|1/2/3} & \pi^f_{123|12/3} & \pi^f_{123|13/2} & \pi^f_{123|1/23} & \pi^f_{123|123} \end{pmatrix}.$$

5. EM ESTIMATION

The EM algorithm can be used to fit the parameters of a mixture model via maximum likelihood estimation (see Dempster, Laird, and Rubin 1977; McLachlan and Peel 2000, p. 47) and has been applied to record linkage problems (e.g., Winkler 1988; Jaro 1989; Larsen and Rubin 2001). Following the model presented in Section 4, let us find the equations of an EM algorithm to estimate Φ. First, for the Expectation step, let us find the conditional distribution of g^j

$$\begin{aligned} P(g^j|\gamma^j) &= \frac{P(x^j)}{P(\gamma^j)} \\ &= \prod_{p\in\mathbb{P}_K} \left[\frac{P(\gamma^j|S_p)P(S_p)}{P(\gamma^j)} \right]^{g_p^j} \\ &= \prod_{p\in\mathbb{P}_K} [P(S_p|\gamma^j)]^{g_p^j}, \end{aligned}$$

that is, $g^j|\gamma^j \sim \text{Multinomial}\,(1, P|\gamma^j)$, where $P|\gamma^j = (P(S_{1/2/\dots/K}|\gamma^j), \dots, P(S_{12\dots K}|\gamma^j))$.

Thus, using the estimation $\hat{\Phi}$ from a previous M step of the algorithm, for the E step, the expectation of the unknown part of g^j is composed by

$$\hat{P}(S_p|\gamma^j) = \frac{\hat{\mathscr{s}}_p \prod_{f=1}^{F} \prod_{p'\in\mathbb{P}_K} \left(\hat{\pi}^f_{p'|p}\right)^{\gamma_{p'}^{j_f}}}{\hat{P}(\gamma^j)} \tag{5}$$

for $p \preccurlyeq p_{b_j}$, where p_{b_j} represents the blocking pattern for r_j. The term $\hat{P}(\gamma^j)$ above is given by

$$\hat{P}(\gamma^j) = \sum_{p\preccurlyeq p_{b_j}} \hat{\mathscr{s}}_p \prod_{f=1}^{F} \prod_{p'\in\mathbb{P}_K} \left(\hat{\pi}^f_{p'|p}\right)^{\gamma_{p'}^{j_f}}.$$

Let $\tilde{g}^j$ be equal to g^j for the entries that are known to be zeroes, and let the remaining entries of $\tilde{g}^j$ be filled with the values given in Equation (5). For the Maximization step, we replace g^j with $\tilde{g}^j$ in the log-likelihood L and estimate Φ via maximum likelihood. We obtain for $\hat{\Pi}$

$$\hat{\pi}^f_{p'|p} = \frac{\sum_{j=1}^{(B_K)^F} n_{\gamma^j} \gamma_{p'}^{j_f} \tilde{g}_p^j}{\sum_{j=1}^{(B_K)^F} n_{\gamma^j} \tilde{g}_p^j},$$

and for $\hat{\mathbf{s}}$ we obtain

$$\hat{\mathscr{s}}_p = \frac{\sum_{j=1}^{(B_K)^F} n_{\gamma^j} \tilde{g}_p^j}{n},$$

where n_{γ^j} represents the frequency counts of each pattern γ^j, as in Jaro (1989). Note that in this case we have $(B_K)^F$ different patterns of γ^j. As usual, the algorithm stops when the values of $\hat{\Phi}$ converge, which can be assessed measuring the distance between $\hat{\Phi}$ in two consecutive iterations. To start this algorithm, we choose initial values taking into account the fact that some probabilities must be greater than others.

5.1 Starting Values

Note that the parameters in each Π^f should hold certain restrictions. In record linkage these constraints are taken into account to start the EM algorithm (Winkler 1993; Lahiri and Larsen 2005). For instance, it is clear that $\pi^f_{12\dots K|12\dots K}$ should be greater than $\pi^f_{1/2/\dots/K|12\dots K}$, that is, given that in a record K-tuple all the entries refer to the same individual, the probability that their information agree should be larger than the probability that all their information disagree. However, note that $\pi^f_{1/2/\dots/K|1/2/\dots/K}$ should not necessarily be greater than $\pi^f_{12\dots K|1/2/\dots/K}$, that is, for a record K-tuple in which all the entries refer to different individuals, the probability that all their information disagree is not necessarily larger than the

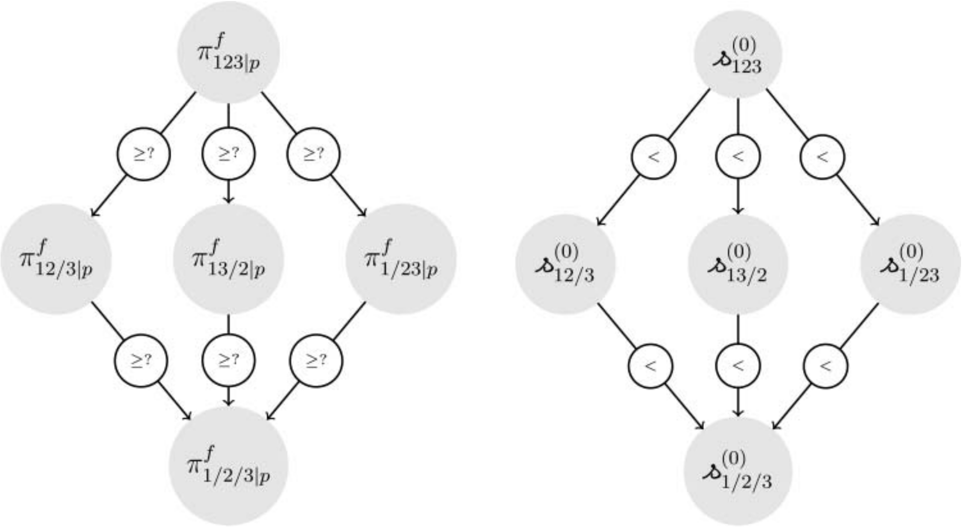


Figure 3. Hasse diagram to determine the set of inequalities between probabilities $\pi^f_{p'|p}$ and $\mathscr{s}^{(0)}_p$. The possible inequalities are established from sources to targets in the arrows, for example, $\mathscr{s}^{(0)}_{123} < \mathscr{s}^{(0)}_{12/3}$.

probability that all their information agree (this is the case for a field with a very common value).

Thus, given the high number of parameters it is not easy to determine which constraints should be taken into account. To determine the set of constraints to start the algorithm, we present a method that uses the natural partial order in $\mathbb{P}_K$. Remember that we say $p' \preccurlyeq p$ if p' is a partition finer than or equal to p. To determine if $\pi^f_{p'|p}$ should be greater or lower than $\pi^f_{p''|p}$ for $p, p', p'' \in \mathbb{P}_K$, we fix the partition p and for all partitions p', p'', such that $p'' \preccurlyeq p' \preccurlyeq p$ we set $\pi^f_{p''|p} \leq \pi^f_{p'|p}$. In any other case we do not have a criterion to order $\pi^f_{p'|p}$ with respect to $\pi^f_{p''|p}$.

Note that this procedure can be visualized using a directed graph in the following way:

1. Construct the Hasse diagram of the partitions $p' \in \mathbb{P}_K$ writing in each node $\pi^f_{p'|p}$, where p is a generic partition.
2. Assign a specific partition to the generic p.
3. Search for the node where $p' = p$.
4. For all the branches under this node, set an inequality $\geq$ between each "father" node and each "son" node.
5. Repeat Steps 2–4 until exhausting the possible partitions.

We can use similar ideas to identify the constraints for $\mathscr{s}_{1/2/\ldots/K}, \ldots, \mathscr{s}_{12\ldots K}$. We simply have that $\mathscr{s}_{p'} \geq \mathscr{s}_p$ whenever $p' \preccurlyeq p$. Naturally, the set of inequalities among the probabilities $\mathscr{s}_p$ can also be represented in a Hasse diagram. Furthermore, if the datafiles being linked have no duplicates, the size of the complete links set $S_{12\ldots K}$ should be smaller than or equal to the smaller datafile size, from which is reasonable to take starting values for $\mathscr{s}_{12\ldots K}$ smaller than $\min\{m_k; k = 1, \ldots, K\}/n$, where m_k represents the number of records in datafile k. In general, we can determine the maximum size of any set S_p if we assume no duplicates into each datafile. Denote q_p as a generic element of the partition $p \in \mathbb{P}_K$, that is, q_p is a subset of $\mathbb{N}_K$. Thus, the maximum size of S_p is $\prod_{q_p \in p} \min\{m_k; k \in q_p\}$, from which is reasonable to start the algorithm taking values lower than $\prod_{q_p \in p} \min\{m_k; k \in q_p\}/n$ for a generic $\mathscr{s}_p$. The starting value for $\mathscr{s}_{1/2/\ldots/K}$ is determined as one minus the other $\mathscr{s}_p$. Notice that since duplicates are rather common in practice, the above values are merely a guide to start the EM algorithm. Finally, since latent class models have multiple solutions corresponding to local maxima of the marginal likelihood, in practice we would take different starting values holding the above constraints, and we would choose the parameters with the maximum marginal likelihood for the observed data γ^j (see, e.g., McLachlan and Peel 2000).

Example 5. We illustrate this procedure for $K = 3$ using the left-hand side of Figure 3. Go to the left panel of Figure 3 and replace p with 123. Since $\pi^f_{123|123}$ is in the top of the graph, we take the set of constraints $\pi^f_{123|123} \geq \pi^f_{12/3|123} \geq \pi^f_{1/2/3|123}$; $\pi^f_{123|123} \geq \pi^f_{13/2|123} \geq \pi^f_{1/2/3|123}$; $\pi^f_{123|123} \geq \pi^f_{1/23|123} \geq \pi^f_{1/2/3|123}$, which correspond to the three different branches under $\pi^f_{123|123}$. Now, replace p with $12/3$. Since in this case the node $\pi^f_{12/3|12/3}$ has only one descendent, we only get the constraint $\pi^f_{12/3|12/3} \geq \pi^f_{1/2/3|12/3}$. This step is similar for $13/2$ and $1/23$. Finally, if we replace p with $1/2/3$, we can see that the node $\pi^f_{1/2/3|1/2/3}$ does not have descendants, so we do not set constraints for the probabilities $\pi^f_{p'|1/2/3}$, $p' \in \mathbb{P}_3$.

For $K = 3$, the right-hand side of Figure 3 represents the set of inequalities for the starting values $\mathscr{s}^{(0)}_p$. We obtain for instance $\mathscr{s}^{(0)}_{1/2/3} > \mathscr{s}^{(0)}_{1/23} > \mathscr{s}^{(0)}_{123}$. Also, for this particular case we take $\mathscr{s}^{(0)}_{123} < \min\{m_1, m_2, m_3\}/n$, $\mathscr{s}^{(0)}_{1/23} < m_1 \min\{m_2, m_3\}/n$, and similar inequalities for $\mathscr{s}^{(0)}_{13/2}$ and $\mathscr{s}^{(0)}_{12/3}$, whereas $\mathscr{s}^{(0)}_{1/2/3} = 1 - \mathscr{s}^{(0)}_{1/23} - \mathscr{s}^{(0)}_{13/2} - \mathscr{s}^{(0)}_{12/3} - \mathscr{s}^{(0)}_{123}$.

6. LINKAGE ASSIGNMENT: GENERALIZED FELLEGI–SUNTER DECISION RULE

The goal of multiple record linkage is to classify each record K-tuple to the appropriate subset S_p. For bipartite record linkage, Fellegi and Sunter (1969) proposed the computation of likelihood ratios as weights for the assignment of record pairs as matched or unmatched pairs. Their procedure is equivalent to test the hypothesis that each record pair belongs to the subset of unmatched record pairs, against the hypothesis that it belongs to the subset of matched pairs, and vice versa.

6.1 Likelihood Ratios and Weights

In multiple record linkage, there are several subsets of records denoting all the possibilities of matching between records from different datafiles. Following Fellegi and Sunter's idea, for each record K-tuple and for each subset, we propose to compute weights following a hypothesis test, where the null hypothesis is the record K-tuple membership to a certain subset, that is, $r_j \in S_p$, against the hypothesis that this record K-tuple does not belong to the subset, that is, $r_j \in S_p^c$, where the superscript c denotes the complement of the set. By using a log-likelihood ratio, we obtain

$$w_p^j = \log \frac{P(\gamma^j | S_p)}{P(\gamma^j | S_p^c)}.$$

The informal idea of the use of the weights w_p^j is that we would order the record K-tuples according to their respective weights and we would assign K-tuples with large w_p^j to the subset S_p. However, the ordering obtained from w_p^j can be obtained in a simpler way, regardless of the model for $P(\gamma^j | S_p)$.

Proposition 1. The ordering obtained from w_p^j, logit$[P(S_p|\gamma^j)]$, and $P(S_p|\gamma^j)$ is the same.

Thus, for ordering and decision purposes we can simply use $P(S_p|\gamma^j)$ (see proofs in the appendix). We still need to determine, however, the cutoffs from which we declare record K-tuples' memberships.

6.2 Cutoff Values

In bipartite record linkage, to declare a record pair as matched or unmatched, the Fellegi–Sunter method orders the possible values of γ^j by their weights in nonincreasing order, determines two cutoff values of the weights, and, according to them, declares matches and nonmatches. For multiple record linkage, we extend this procedure and prove its optimality.

Theorem 1. The decision procedure described below maximizes the probability of assigning each record K-tuple to the right subset, subject to a set of admissible error levels μ_p.

1. Each record K-tuple is *potentially* declared to belong to the subset S_p if and only if p is the pattern for which $P(S_p|\gamma^j)$ is maximum among all possible patterns in $\mathbb{P}_K$. Thus, the set of record K-tuples is partitioned into B_K subsets, and for each record K-tuple in one of these partitions we consider only two possibilities, whether to declare it to belong to the subset S_p or to keep it undeclared.
2. For the record K-tuples in each partition, we order the possible values of γ^j by their weights (or equivalently by $P(S_p|\gamma^j)$) in nonincreasing order indexing by the subscript $(j)_p$.
3. We find one value $(j')_p$ for each set of weights related to each subset, to determine the record K-tuple memberships. The value $(j')_p$ is found such that
$$\mu_p = \sum_{(j)_p=1}^{(j')_p-1} P\left(\gamma^{(j)_p} | S_p^c\right),$$
where $\mu_p = P(\text{assign } r_j \text{ the membership of } S_p | r_j \in S_p^c)$ is an admissible error level. Each $P(\gamma^{(j)} | S_p^c)$ can be computed as
$$P\left(\gamma^{(j)_p} | S_p^c\right) = \frac{\sum_{p' \in \mathbb{P}_K, p' \neq p} P\left(\gamma^{(j)_p} | S_{p'}\right) \mathscr{s}_{p'}}{1 - \mathscr{s}_p}.$$
4. Finally, for those record K-tuples with configurations of $\gamma^{(j)_p}$, $(j)_p = 1, \ldots, (j')_p - 1$, we decide that they belong to the subset S_p. For those record K-tuples with configurations $\gamma^{(j)_p}$ with $(j)_p \geq (j')_p$, we keep them undeclared.

In the Appendix, we show that the above decision rule is optimal under the availability of the true matching probabilities. We show that this decision rule minimizes the probability of assigning each record K-tuple to the wrong subset S_p or keeping it undeclared, subject to a set of admissible error levels μ_p, or namely, it maximizes the probability of assigning each record K-tuple to the right subset, subject to a set of admissible error levels μ_p. The Fellegi–Sunter decision rule for bipartite record linkage can be obtained as a corollary of Theorem 1. In practice the optimality of this decision rule depends on the quality of the estimation of the matching probabilities. Belin and Rubin (1995) and Larsen and Rubin (2001) provided evidence that nominal and actual error levels disagree in different applications. Belin and Rubin (1995) proposed a method to calibrate error rates as a function of cutoff values for bipartite record linkage. This is an important problem that we expect to address in our ongoing work for the multiple record linkage context.

7. LINKING HOMICIDE RECORD SYSTEMS IN COLOMBIA

The Colombian homicide data described in Section 1.2 were provided by the Conflict Analysis Resource Center (CERAC) where a linkage by hand was performed for a subset of the data, corresponding to the province of Quindio for the last three months of 2004. In this section, we present an application to the integration of these three datafiles. In this period, 67, 62, and 33 homicides were recorded by the Census Bureau, the National Police, and the Forensics Institute, respectively. The common fields of these three datafiles are town and date of the homicide, gender, and age of the victim.

An outline of the implementation of the method is as follows:

1. Find the set of record triplets that are suitable for classification into the different matching patterns. This set is obtained after blocking.
2. Compute the comparison data according to the possible patterns of agreement for all the triplets to be classified and for every common field.
3. Train the mixture model of the distribution of the comparison data.
4. Divide the set of triplets according to the subsets S_p for which $\hat{P}(S_p|\gamma^j)$ is maximum.

Table 2. Error rates of multiple record linkage assignments for Census (1)–Forensics (2)–Police (3) record triplets. Three comparison data options for age of the victim and date of the homicide. OME: overall misclassification error, MWGE: mean within group error

		Misclassification error						
	Age and date data	1/2/3	12/3	13/2	1/23	123	OME	MWGE
1.	Exact comparisons	0.6203	0.2216	0.3915	0.0079	0.4444	0.5977	0.3371
2.	Three comparison categories	0.0470	0.0109	0.0803	0.0510	0.0370	0.0471	0.0453
3.	Three comparison categories + blocking Kid–Young–Elderly	0.0365	0.0079	0.0598	0.0082	0.0370	0.0359	0.0299

5. Within each subset, sort the triplets by $\hat{P}(S_p|\gamma^j)$ and use an admissible error level to either declare the triplets as belonging to the subset S_p or keep them undeclared.

To implement the method, we used town of the homicide and gender of the victim for blocking. We assigned the membership to the subset $S_{1/2/3}$ to the triplets with blocking pattern 1/2/3. We used the proposed method to classify the remaining triplets. To use date of the homicide and age of the victim, we explored several options, but we only report the results of using three of them (Table 2). The first option only includes exact comparison data for both variables. The second option constructs three categorical variables from each variable age and date, and creates comparison data using these new categorical variables. These variables are constructed in the following fashion: the categories of the variable AgeA are 0–2, 3–5, and so on; the categories of the variable AgeB are 0, 1–3, 4–6, and so on; and finally, the categories of the variable AgeC are 0–1, 2–4, and so on. A similar procedure is used for date of the homicide, starting from the first day of the period of the data. The third approach uses the previous categorical variables and in addition exploits a specific structure of the age recorded in these datasets to create an additional blocking variable. The ages recorded in these three datafiles present two gaps, this is, there are no homicides recorded in the 5–11 and 56–65 age intervals. Thus, we create a new blocking variable that classifies "kids," "young," and "elderly" individuals. We think it is safe to use this variable for blocking since no records with similar ages are assigned to different blocks. Also, to help the EM algorithm to identify the appropriate clusters, we replaced $\hat{P}(S_p|\gamma^j)$ by 1 for those triplets with $\gamma_p^{j_f} = 1$ for all the fields f and for $p \in \{12/3, 13/2, 1/23, 123\}$. This semisupervised approach is a missing data problem under multinomial sampling (Dempster, Laird, and Rubin 1977). We made the final assignments using nominal error levels $\mu_p = 0.01$ for all p.

In Table 2, we present different measures of the performance of the multiple record linkage decisions using the three different options for the inclusion of the information about age of the victim and date of the homicide. These measures were obtained after comparing with the results of the hand matching procedure, which is thought to be more reliable. Besides the usual misclassification errors, we present the mean within group error (MWGE) rate (Qiao and Liu 2009), which controls the different sizes of the clusters S_p by taking the average of the error rates for each S_p. From the first age and date comparison data, we can see that the multiple record linkage procedure can produce catastrophic results if it is not used carefully. For this scenario all the misclassification errors are very high, which indicates that the multiple record linkage process did not find the appropriate clusters. For the first comparison data, only exact comparisons were included, hence small differences in age and date were treated the same as large differences. For the second age and date comparison data, the results improved significantly. The way these comparison variables were created is such that if there is exact agreement in age or in date, the three corresponding comparison variables agree. If there is a difference of one unit, two of them agree, and if there is a difference of two units, only one of the variables agree. This approach is more flexible to capture small measurement error in age and date. The final approach additionally blocks three categories of age, which helps to reduce the number of misclassified triplets. For this final approach, all the measures of misclassification error are very close to zero, which indicates that multiple record linkage can provide good results if used properly. Naturally, the good performance of the method depends on the specific datafiles to be linked and the models implemented.

We performed a bipartite record linkage for each of the three pairs of datafiles using the same blocking variables and the same comparison data as the third approach in Table 2. The assignments were also made using nominal error levels of 0.01. For the triplets on which a decision could be made, the overall misclassification error was 0.0435 and the MWGE was 0.0311. When trying to combine the decisions of the three independent procedures, however, we obtained a set of 43 record triplets on which we could not assign a decision. Among this set of record triplets, the multiple record linkage procedure coincided with the hand matching procedure in 32 cases (74%). Of course the performance of the method for those record triplets is not as good as the general performance, since these record triplets are usually the ones that are more difficult to classify. However, multiple record linkage provides a decision along with a measure of uncertainty for that decision (namely, the matching probabilities), something that is not available from reconciling bipartite record linkages.

8. SIMULATION STUDIES

In practice, the performance of our method will depend on several factors: (a) the amount of measurement error of the datafiles, (b) the number of common variables and their number of categories/variability, (c) the sizes of the datafiles and their overlaps, (d) the dependence structure among the recorded fields, and (e) the existence of replicate records in the datafiles, etc. Here, we explore the performance of the proposed method under some simple scenarios, emphasizing how measurement error affects our results. We used the R language to perform our simulations (R Development Core Team 2010).

8.1 Generating Measurement Error

Tancredi and Liseo (2011) used a simplified version of the *hit-miss* model (Copas and Hilton 1990) to generate measurement error. This model for categorical information on records measured with error is given by

$$P\big(Y_f^{\text{obs}} = y_f^{c^o} | Y_f = y_f^c\big) = (1-\beta_f)I\big(y_f^{c^o} = y_f^c\big) + \beta_f / C_f, \qquad (6)$$

where Y_f^{obs} represents the observed field f and Y_f represents the true value of the field f. Both Y_f^{obs} and Y_f have support $\{y_f^1, \dots, y_f^c, \dots, y_f^{C_f}\}$, where C_f represents the number of categories of the field f. Equation (6) includes a measurement error parameter β_f that represents the probability of measurement error for the field f. This model establishes that conditioning on the unobserved true values, we can model each single record field as a mixture of two components: the first component is concentrated on the true value while the second one is uniformly distributed over the support of the field (Tancredi and Liseo 2011). In our simulation studies, we do not generate error for the blocking variable. For the numerical variables, we generate error using the following model:

$$\begin{aligned} P\big(Y_f^{\text{obs}} = y_f^{c^o} | Y_f = y_f^c\big) &= (1-\beta_f)I\big(y_f^{c^o} = y_f^c\big) \\ &\quad + \beta_f \frac{2}{5} 2^{-|y_f^{c^o} - y_f^c|} I\big(|y_f^{c^o} - y_f^c| < 3\big), \end{aligned} \qquad (7)$$

which allows measurement error around the true value. For our simulation study, we consider the same value of β_f for all the fields subject to error (so we drop the subindex f).

8.2 To Block or Not to Block?

Blocking is usually an important component of record linkage since working with the complete Cartesian product of the datafiles is computationally inefficient. In this section, we show that we need blocking to obtain good classification results. Thus, we may want to block even in the presence of adequate computational power to handle the record linkage process on the complete Cartesian product.

We take the Census homicide data as the true population information and we generate three equal size datafiles subject to

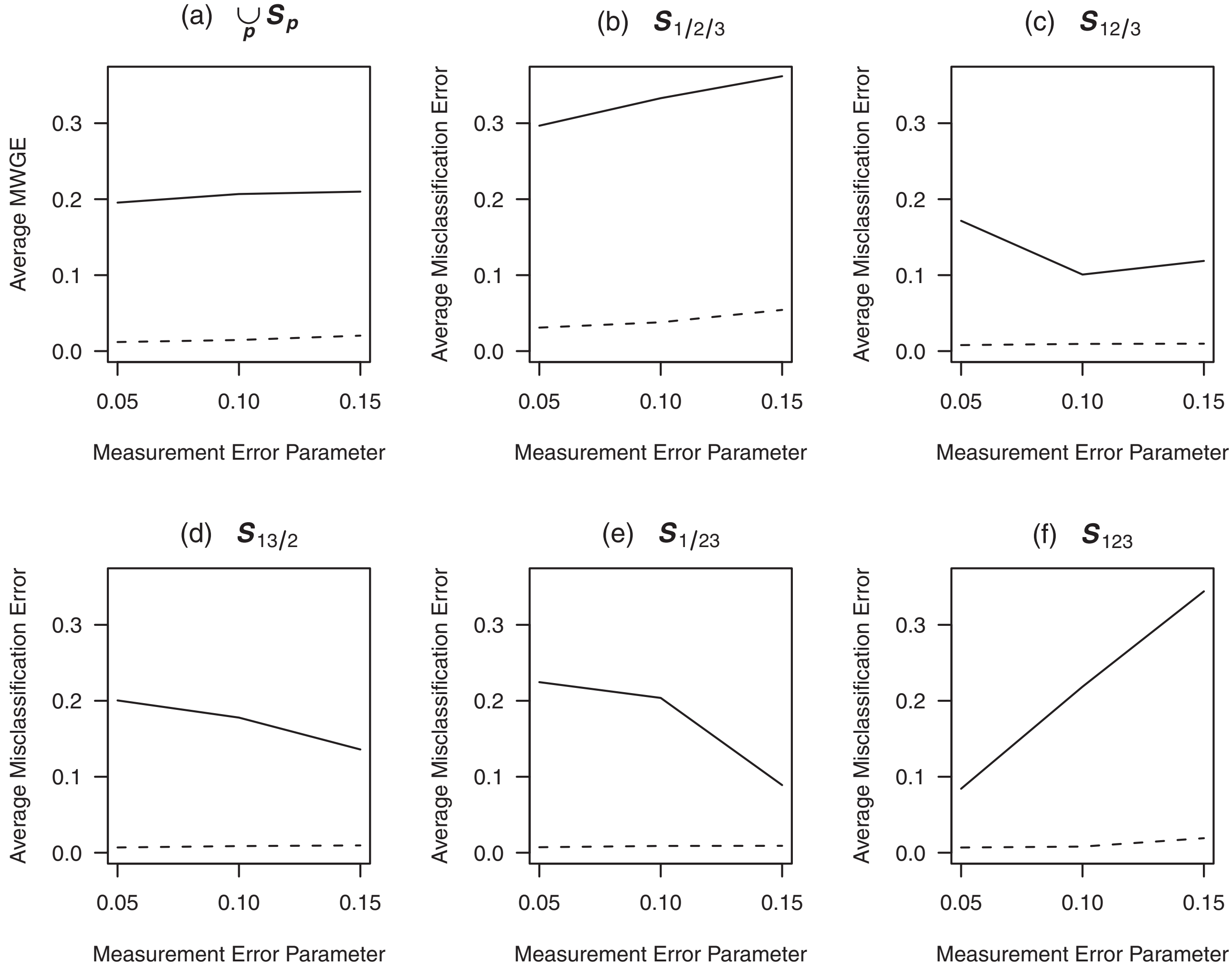


Figure 4. Measures of misclassification error for nonblocking (solid line) and blocking (dashed line) scenarios.

measurement error. We generate measurement error according to the model (7) for date of the homicide and age of the victim. We do not generate measurement error for sex of the victim and city of the homicide since we use these variables for blocking. We simulate 100 triplets of datafiles and for each triplet, we perform multiple record linkage using the second option of comparison data presented in Section 7. In Figure 4, we present the performance results for three values of the measurement error parameter: 0.05, 0.10, and 0.15. We compare the results of our method without blocking (solid line) and after blocking by gender of the victim and city of the homicide (dashed line). In panel (a) of Figure 4, we average over all the simulations the MWGE as a measure of the general performance of the method (or in other words, a measure of the performance of our method on $\bigcup_p S_p$). In panels (b)–(f), we present the average misclassification error for each specific subset S_p.

We can see that, for this example, the effect of blocking is huge. In general, the error rates are very large when we use no blocking, but they decay to values close to zero under blocking. Note also that the larger the measurement error, the larger the error recovering the subsets S_{123} and $S_{1/2/3}$, which indicates that measurement error causes true triple links to be missed and false links to be created.

8.3 Number of Blocks and Low-Quality Fields

In certain applications, there are different blocking options and the possibility to include low-quality fields in the linkage process. In this section, we explore these scenarios. We generate three databases containing five independent common fields across the different scenarios. These first five fields contain 3, 5, 10, 10, and 15 categories, respectively, and each category is generated with equal probability. We also use one additional

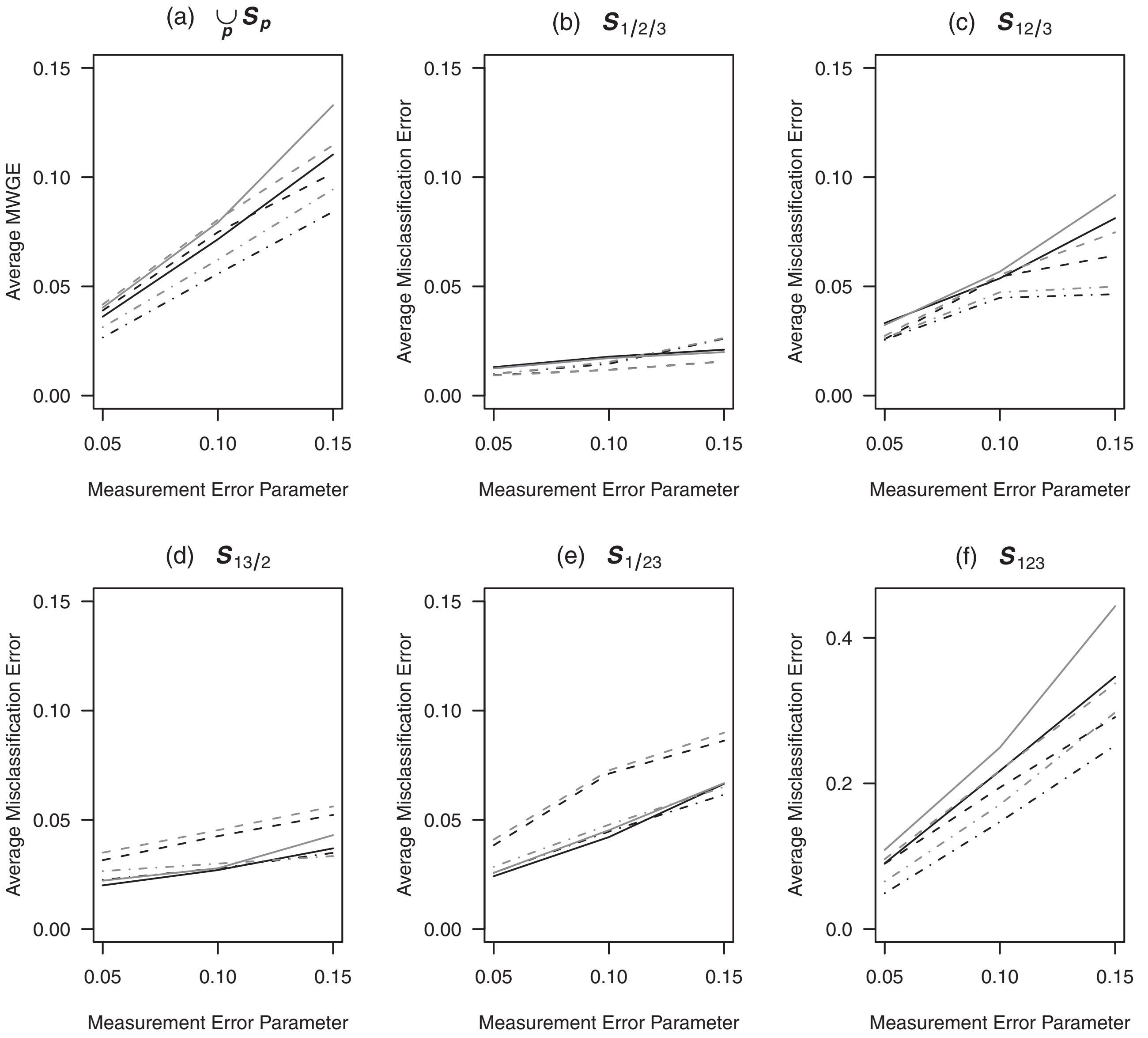


Figure 5. Measures of misclassification error for different number of blocks and inclusion/exclusion of low-quality fields. The blocking scenarios are 5 blocks (solid line), 10 blocks (dashed line), and 15 blocks (dot-dashed line). The gray lines represent the performance of the method including the low-quality extra field. Note the different scale of panel (f).

independent blocking variable to check the performance of the method under blocking. We consider three different blocking scenarios that correspond to 5, 10, and 15 categories of the blocking variable, where the categories are generated with equal probability. For all the simulation scenarios, the sizes of the databases and their overlaps are the same as in the Colombian homicide data.

For one of the fields with 10 categories, we use $\beta = 0.7$ to simulate a scenario where a common variable is available, but it is known that its quality is low. We keep $\beta = 0.7$ for the previous variable across three different measurement error scenarios for the remaining four fields. These three scenarios correspond to three different values of β: 0.05, 0.10, and 0.15, and in each scenario the same β is used to generate error for the remaining four fields. Given the three true databases, we generate 100 triplets of observed databases using the hit-miss model (6). For each triplet of databases, we performed six implementations of the proposed methodology for multiple record linkage. The six implementations correspond to the combination of including/excluding the low-quality field and the three blocking options. We made the final assignments using nominal error levels $\mu_p = 0.01$ for all p.

To evaluate the performance of the method in terms of recovering the classes S_p, we report the misclassification error rate for each class S_p and the MWGE rate (Qiao and Liu 2009) for the triplets that were assigned to a certain group. The MWGE rate is more meaningful than the overall misclassification error for record linkage since the groups S_p are extremely unbalanced, for example, the subset $S_{1/2/3}$ is massive whereas the subset S_{123} is extremely small. We present the results in Figure 5, where panel (a) shows the average over all the simulations of the MWGE and panels (b)–(f) show the average misclassification error for each class S_p. All the panels show the performance measures as a function of the measurement error parameter. The solid, dashed, and dot-dashed lines represent the error values for the method with 5, 10, and 15 blocks, respectively. The gray lines represent the method including the low-quality extra field. Note that the scale of the vertical axes is the same for panels (a)–(e), but we present panel (f) with a different scale since the errors for the subset S_{123} are significantly larger compared with the other subsets.

We can see that, in general, the larger the measurement error, the larger the error rates, which is something that one would expect. We can also see that under all the scenarios, increasing the amount of blocking decreases the error rates. In particular, note in panel (f) that blocking has a huge impact on the reduction of the misclassification for the class S_{123}. Finally, we note that for each blocking scenario, the inclusion of the low-quality extra field increases the error rates.

9. CONCLUSIONS AND FUTURE WORK

Our method provides a framework for the integration of more than two datafiles without common identifiers. The ideas are an extension of the theory proposed by Fellegi and Sunter (1969) and its more modern implementations, as in Winkler (1988) and Jaro (1989). The method solves the problem of obtaining nontransitive decisions, as it is common when reconciling bipartite record linkages. Our method also provides matching probabilities for the record K-tuples, something that is not available from reconciling bipartite record linkages, but that is necessary to incorporate the uncertainty of the linkage procedure in posterior analysis such as regression (Lahiri and Larsen 2005). We proposed a decision rule that is optimal under the availability of the true matching probabilities. In practice, however, the optimality of the decision rule hinges on the availability of well-calibrated probability models, that is, good estimates of the probability of a particular K-tuple belonging to the subsets S_p. Thus, we need to consider models that go beyond the present one and that capture dependencies between fields (see, e.g., Larsen and Rubin 2001). Nevertheless, even using a naive model, our method performed well both in the integration of the Colombian homicide datafiles and in our simulations.

We believe our method holds promise in the context of record linkage for census coverage measurement evaluation. For example, the U.S. Census Bureau has for several decades done a two-sample linkage between the actual enumeration and data from a post-enumeration survey based on data from a nationwide sample of census blocks (Hogan 1992, 1993). Additional sources of data that could be used to improve coverage estimation include the American Community Survey and various administrative record files. Incorporation of them would require linkage of $K \geq 3$ datafiles, using methods that could build upon the work described here that would take into account multiple sampling designs and census adjustments such as imputations and erroneous enumerations.

APPENDIX: PROOFS

In the proofs presented below, we use the notation introduced in Section 4, where for instance, $P(S_p^c)$ means $P(r_j \in S_p^c)$, and so on. Also, for functions $f(x)$ and $g(x)$ we denote $f(x) \propto g(x)$ if $f(x) = g(x) + t$, for some constant t.

Proof of Proposition 1. The ordering of w_p^j is the same as the ordering of $\text{logit}[P(S_p|\gamma^j)]$ since

$$\begin{aligned} w_p^j &= \log \frac{P(S_p|\gamma^j)/P(S_p)}{P\left(S_p^c|\gamma^j\right)/P\left(S_p^c\right)} \\ &\propto \log \frac{P(S_p|\gamma^j)}{P\left(S_p^c|\gamma^j\right)} \\ &= \text{logit}[P(S_p|\gamma^j)]. \end{aligned}$$

Finally, the logit function is a monotonic increasing function of its argument, thus the ordering of $\text{logit}[P(S_p|\gamma^j)]$ is the same as the ordering of $P(S_p|\gamma^j)$.

Proof of Theorem 1. Optimality of the Generalized Fellegi–Sunter Linkage Rule.

Let us define the set of possible decisions for a record K-tuple. Let us call D_p the decision of assigning a record K-tuple to the subset S_p and D_u the decision to keep the record K-tuple undeclared. Thus, a decision function d is a $(B_K + 1)$-tuple given by

$$\begin{aligned} d(\gamma^j) = (&P(D_{1/2/\ldots/K}|\gamma^j), \ldots, P(D_p|\gamma^j), \ldots, P(D_{12\ldots K}|\gamma^j), \\ &P(D_u|\gamma^j)), \end{aligned}$$

where

$$P(D_u|\gamma^j) + \sum_{p \in \mathbb{P}_K} P(D_p|\gamma^j) = 1.$$

The proposed decision rule L_0 is such that

$$\begin{aligned} P_0(D_p|\gamma^j) = 1, &\quad \text{if } (j)_p \le (j')_p - 1; \\ P_0(D_u|\gamma^j) = 1, &\quad \text{if } (j)_p \ge (j')_p; \end{aligned}$$

for $(j)_p$ in the subset of record K-tuples for which $P(S_p|\gamma^j)$ is maximum and $(j')_p$ is obtained as in the statement of Theorem 1. This decision rule minimizes the probability of assigning each record K-tuple to the wrong subset S_p or keeping it undeclared, subject to a set of admissible error levels $\mu_p = P(D_p|S_p^c)$, $p \in \mathbb{P}_K$. For decision rules L_0 and L_1

$$\begin{aligned} \mu_p = P\left(D_p|S_p^c\right) &= \sum_{(j)_p} P_0\left(D_p|\gamma^{(j)_p}\right) P\left(\gamma^{(j)_p}|S_p^c\right) \\ &= \sum_{(j)_p} P_1\left(D_p|\gamma^{(j)_p}\right) P\left(\gamma^{(j)_p}|S_p^c\right). \end{aligned}$$

From the construction of L_0, we obtain

$$\sum_{(j)_p \le (j')_p - 1} P\left(\gamma^{(j)_p}|S_p^c\right) = \sum_{(j)_p} P_1\left(D_p|\gamma^{(j)_p}\right) P\left(\gamma^{(j)_p}|S_p^c\right)$$

or

$$\begin{aligned} &\sum_{(j)_p \le (j')_p - 1} P\left(\gamma^{(j)_p}|S_p^c\right)\left[1 - P_1\left(D_p|\gamma^{(j)_p}\right)\right] \\ &\quad = \sum_{(j)_p \ge (j')_p} P_1\left(D_p|\gamma^{(j)_p}\right) P\left(\gamma^{(j)_p}|S_p^c\right). \end{aligned} \tag{A.1}$$

Since

$$P\left(\gamma^{(i)_p}|S_p\right) P\left(\gamma^{(j)_p}|S_p^c\right) \le P\left(\gamma^{(j)_p}|S_p\right) P\left(\gamma^{(i)_p}|S_p^c\right)$$

whenever $(j)_p < (i)_p$, we have

$$\begin{aligned} &\left[\sum_{(j)_p \ge (j')_p} P_1\left(D_p|\gamma^{(j)_p}\right) P\left(\gamma^{(j)_p}|S_p\right)\right] \\ &\quad \times \left[\sum_{(j)_p \le (j')_p - 1} P\left(\gamma^{(j)_p}|S_p^c\right)\left[1 - P_1\left(D_p|\gamma^{(j)_p}\right)\right]\right] \\ &\quad \le \left[\sum_{(j)_p \ge (j')_p} P_1\left(D_p|\gamma^{(j)_p}\right) P\left(\gamma^{(j)_p}|S_p^c\right)\right] \\ &\quad \times \left[\sum_{(j)_p \le (j')_p - 1} P\left(\gamma^{(j)_p}|S_p\right)\left[1 - P_1\left(D_p|\gamma^{(j)_p}\right)\right]\right]; \end{aligned} \tag{A.2}$$

dividing (A.2) by (A.1), we obtain

$$\begin{aligned} &\left[\sum_{(j)_p \ge (j')_p} P\left(\gamma^{(j)_p}|S_p\right) P_1\left(D_p|\gamma^{(j)_p}\right)\right] \\ &\quad \le \left[\sum_{(j)_p \le (j')_p - 1} P\left(\gamma^{(j)_p}|S_p\right)\left[1 - P_1\left(D_p|\gamma^{(j)_p}\right)\right]\right] \end{aligned}$$

from which

$$\left[\sum_{(j)_p} P\left(\gamma^{(j)_p}|S_p\right) P_1\left(D_p|\gamma^{(j)_p}\right)\right] \le \left[\sum_{(j)_p} P\left(\gamma^{(j)_p}|S_p\right) P_0\left(D_p|\gamma^{(j)_p}\right)\right],$$

which is the same as

$$P_1(D_p|S_p) \le P_0(D_p|S_p),$$

which implies

$$P_1\left(D_p^c|S_p\right) \ge P_0\left(D_p^c|S_p\right) \tag{A.3}$$

for all $p \in \mathbb{P}_K$. Note that the probability of taking a wrong decision or not deciding can be written as

$$\sum_{p \in \mathbb{P}_K} P\left(D_p^c \cap S_p\right) = \sum_{p \in \mathbb{P}_K} P\left(D_p^c|S_p\right) P(S_p),$$

which is minimized by the generalized Fellegi–Sunter linkage rule L_0, as we can see using (A.3).

[Received April 2012. Revised September 2012.]

REFERENCES

Anderson, M. J., and Fienberg, S. E. (1999), *Who Counts?: The Politics of Census-Taking in Contemporary America*, New York: Russell Sage Foundation (revised paperback (2001) edition). [386]

Asher, J., and Fienberg, S. E. (2001), "Statistical Variations on an Administrative Records Census," in *Proceedings of the Section on Government Statistics of the American Statistical Association, American Statistical Association*. [385]

Asher, J., Fienberg, S. E., Stuart, E., and Zaslavsky, A. (2003), "Inferences for Finite Populations Using Multiple Data Sources with Different Reference Times," in *Proceedings of Statistics Canada Symposium 2002: Modelling Survey Data For Social and Economic Research. Statistics Canada*. [385]

Ball, P. (2000), "The Guatemalan Commission for Historical Clarification: Generating Analytic Reports; Inter-Sample Analysis," in *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*, eds. P. Ball, H. F. Spirer, and L. Spirer, pp. 259–286, Washington, D.C.: American Association for the Advancement of Science. [386]

Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False–Match Rates in Record Linkage," *Journal of the American Statistical Association*, 90, 694–707. [391]

Bell, R. M., Keesey, J., and Richards, T. (1994), "The Urge to Merge: Linking Vital Statistics Records and Medicaid Claims," *Medical Care*, 32, 1004–1018. [385]

Bilenko, M., Mooney, R. J., Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003), "Adaptive Name Matching in Information Integration," *IEEE Intelligent Systems*, 18, 16–23. [385]

Christen, P. (2012), "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication," *IEEE Transactions on Knowledge and Data Engineering*, 24, 1537–1555. [387]

Copas, J. B., and Hilton, F. J. (1990), "Record Linkage: Statistical Models for Matching Computer Records," *Journal of the Royal Statistical Society,* Series A, 153, 287–320. [386,393]

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (1993), "A Three-Sample Multiple-Recapture Approach to Census Population Estimation With Heterogeneous Catchability," *Journal of the American Statistical Association*, 88, 1137–1148. [385]

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. [386,389,392]

Departamento Administrativo Nacional de Estadisticas, DANE. (2009), "Metodologia Estadisticas Vitales," Technical Report 82, Direccion de Censos y Demografia. [386]

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210. [385,386,390,395]

Fienberg, S. E., and Manrique-Vallier, D. (2009), "Integrated Methodology for Multiple Systems Estimation and Record Linkage Using a Missing Data Formulation," *AStA Advances in Statistical Analysis*, 93, 49–60. [385]

Gohdes, A. (2010), "Different Convenience Samples, Different Stories: The Case of Sierra Leone," Technical Report, The Human Rights Data Analysis Group at Benetech. [386]

Herzog, T. H., Scheuren, F., and Winkler, W. E. (2010), "Record Linkage," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 535–543. [385,387]

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007), *Data Quality and Record Linkage Techniques*. New York: Springer. [385,387]

Hogan, H. (1992), "The 1990 Post-Enumeration Survey: An Overview," *The American Statistician*, 46, 261–269. [385,395]

——— (1993), "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association*, 88, 1047–1060. [385,395]

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. [385,386,388,389,395]

Lahiri, P., and Larsen, M. D. (2005), "Regression Analysis With Linked Data," *Journal of the American Statistical Association*, 100, 222–230. [389,395]

Larsen, M. D., and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 96, 32–41. [388,389,391,395]

Lum, K., Price, M., Guberek, T., and Ball, P. (2010), "Measuring Elusive Populations With Bayesian Model Averaging for Multiple Systems Estimation: A Case Study on Lethal Violations in Casanare, 1998–2007," *Statistics, Politics, and Policy*, 1. [386]

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley. [389,390]

Méray, N., Reitsma, J. B., Ravelli, A. C., and Bonsel, G. J. (2007), "Probabilistic Record Linkage is a Valid and Transparent Tool to Combine Databases Without a Patient Identification Number," *Journal of Clinical Epidemiology*, 60, 883–891. [385]

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954–959. [385]

Qiao, X., and Liu, Y. (2009), "Adaptive Weighted Learning for Unbalanced Multicategory Classification," *Biometrics*, 65, 159–168. [392,395]

R Development Core Team. (2010), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [393]

Restrepo, J. A., and Aguirre, K. (2007), *Homicidios y Muertes Violentas: Un Analisis Comparativo de las Fuentes en Colombia*, Forensis Datos Para La Vida, Bogotá, Columbia: Instituto Nacional de Medicina Legal y Ciencias Forenses. [386]

Rosen, K. H. (2006), *Discrete Mathematics and Its Applications* (6th ed.), Boston: McGraw-Hill. [388]

Rota, G.-C. (1964), "The Number of Partitions of a Set," *The American Mathematical Monthly*, 71, 498–504. [386]

Tancredi, A., and Liseo, B. (2011), "A Hierarchical Bayesian Approach to Record Linkage and Size Population Problems," *The Annals of Applied Statistics*, 5, 1553–1585. [393]

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi–Sunter Model of Record Linkage," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 667–671. [385,386,388,389,395]

——— (1993), "Improved Decision Rules in the Fellegi–Sunter Model of Record Linkage," in *Proceedings of Survey Research Methods Section, American Statistical Association*, pp. 274–279. [389]

Winkler, W. E., and Thibaudeau, Y. (1991), "An Application of the Fellegi–Sunter Model of Record Linkage to the 1990 U.S. Decennial Census," Statistical Research Division Technical Report 91-9, U.S. Bureau of the Census. [385]

Zaslavsky, A. M., and Wolfgang, G. S. (1993), "Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative-List Data," *Journal of Business and Economic Statistics*, 11, 279–288. [385]