Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Maximum entropy classification for record linkage

by Danhyang Lee, Li-Chun Zhang and Jae Kwang Kim

Release date: June 21, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© Her Majesty the Queen in Right of Canada as represented by the Minister of Industry, 2022

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 1-23
Statistics Canada, Catalogue No. 12001-X

Maximum entropy classification for record linkage

Danhyang Lee, Li-Chun Zhang and Jae Kwang Kim[1]

Abstract

By record linkage one joins records residing in separate files which are believed to be related to the same entity. In this paper we approach record linkage as a classification problem, and adapt the maximum entropy classification method in machine learning to record linkage, both in the supervised and unsupervised settings of machine learning. The set of links will be chosen according to the associated uncertainty. On the one hand, our framework overcomes some persistent theoretical flaws of the classical approach pioneered by Fellegi and Sunter (1969); on the other hand, the proposed algorithm is fully automatic, unlike the classical approach that generally requires clerical review to resolve the undecided cases.

Key Words: Probabilistic linkage; Density ratio; False link; Missing match; Survey sampling

1. Introduction

Combining information from multiple sources of data is a frequently encountered problem in many disciplines. To combine information from different sources, one assumes that it is possible to identify the records associated with the same entity, which is not always the case in practice. The entity may be individual, company, crime, etc. If the data do not contain unique identification number, identifying records from the same entity becomes a challenging problem. Record linkage is the term describing the process of joining records that are believed to be related to the same entity. While record linkage may entail the linking of records within a single computer file to identify duplicate records, referred to as deduplication, we focus on linking of records across separate files.

Record linkage (RL) has been employed for several decades in survey sampling producing official statistics. In particular, linking administrative files with survey sample data can greatly improve the quality and resolution of the official statistics. As applications, Jaro (1989) and Winkler and Thibaudeau (1991) merged post-enumeration survey and census data for census coverage evaluation. Zhang and Campbell (2012) linked population census data files over time, and Ørnen, Jones and Ralphs (2015) linked administrative registers to create a single statistical population dataset. The classical approach pioneered by Fellegi and Sunter (1969), which is the most popular method of RL in practice, has been successfully employed for these applications.

The probabilistic decision rule of Fellegi and Sunter (1969) is based on the likelihood ratio test idea, by which we can determine how likely a particular record pair is a true match. In applying the likelihood ratio test idea, one needs to estimate the model parameters of the underlying model and determine the thresholds of the decision rule. Winkler (1988) and Jaro (1989) treat the matching status as an unobserved variable and propose an EM algorithm for parameter estimation, which we shall refer to as the

1. Danhyang Lee, Department of Information Systems, Statistics and Management Science, University of Alabama, Tuscaloosa, AL, U.S.A. Li-Chun Zhang, Department of Social Statistics and Demography, University of Southampton, Southampton, U.K., Statistics Norway, Oslo, Norway and Department of Mathematics, University of Oslo, Oslo, Norway. E-mail: L.Zhang@soton.ac.uk. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA, U.S.A.

FS-procedure. See Herzog, Scheuren and Winkler (2007), Christen (2012) and Binette and Steorts (2020) for overviews. However, as explained in Section 2.1, to motivate the FS-procedure as an EM algorithm requires the crucial assumption that measures of agreement between the record pairs, called comparison vectors, are independent from one record pair to another, which is impossible to hold in reality. Newcombe, Kennedy, Axford and James (1959) address dependence between comparison vectors through data application. Also, see e.g. Tancredi and Liseo (2011), Sadinle (2017), and Binette and Steorts (2020) for discussions of this issue. Bayesian approaches to RL are also available in the literature (Steorts, 2015; Sadinle, 2017; Stringham, 2021). Bayesian approaches to RL problems allow us to quantify uncertainty on the matching decisions. However, the stochastic search using MCMC algorithm in the Bayesian approach involves extra computational burden.

To develop an alternative approach, we first note that the RL problem is essentially a classification problem, where each record pair is associated with a latent class indicator, namely match or nonmatch. Thus, classification techniques based on machine learning approaches have been employed for record linkage (Hand and Christen, 2018; Christen, 2012, 2008; Sarawagi and Bhamidipaty, 2002). In this paper, we adapt the maximum entropy method for classification to record linkage. Specifically, we can view the likelihood ratio of the method proposed by Fellegi and Sunter (1969) as a special case of the density ratio and apply the maximum entropy method for density ratio estimation. For example, Nigam, Lafferty and McCallum (1999) use the maximum entropy for text classification and Nguyen, Wainwright and Jordan (2010) develop a more unified theory of maximum entropy method for density ratio estimation. There is, however, a key difference of record linkage to the standard setting of classification problems, in that the latent class indicators, or the match statuses, for record pairs are dependent.

We present our maximum entropy record linkage algorithm for both supervised and unsupervised settings, while our main contributions concern the unsupervised case. Supervised approaches need training data, i.e., record pairs with known true match and true nonmatch status. Such training data are often not available in real world situations, or have to be prepared manually, which is very expensive and time consuming (Christen, 2007). Thus, the unsupervised case is by far the most common in practice. In the unsupervised case, however, one cannot estimate the density ratio directly based on the observed true matches and nonmatches, and it is troublesome to jointly model for the unobserved match statuses and observed comparison scores over all the record pairs. We develop a new iterative algorithm to jointly estimate the density ratio as well as the maximum entropy classification set in the unsupervised setting and prove its convergence. The associated algorithms of the linkage uncertainty are also developed.

Furthermore, we show that the FS-procedure can be incorporated as a special case of our approach to estimation, but without the need of the independence assumption between the record pairs. This reveals that the FS-procedure can be motivated without the independence assumption, and explains why it gives reasonable results in many situations. The choice of the set of links is guided by the uncertainty measures developed in this paper. This is an important practical improvement over the classical approach, which does not directly provide any uncertainty measure for the final set of links. Our procedure is fully

automatic, without the need for resource-demanding clerical review that is required under the classical approach.

The paper is organised as follows. In Section 2, the basic setup and the classical approach are introduced. In Section 3, the proposed method is developed under the setting of supervised record linkage. In Section 4, we extend the proposed method to the more challenging case of the unsupervised record linkage. Discussions of some related estimation approaches and technical details are presented in Section 5 and the supplementary material. Results from an extensive simulation study are presented in Section 6. Some concluding remarks and comments on further works are given in Section 7.

2. Problems with the classical approach

Suppose that we have two data files A and B that are believed to have many common entities but no duplicates within each file. Any record in A and another one in B may or may not refer to the same entity. Our goal is to find the true matches among possible pairs of the two data files. Let the bipartite comparison space $\Omega = A \times B = M \cup U$ consist of matches M and non-matches U between the records in files A and B. For any pair of records $(a, b) \in \Omega$, let γ_{ab} be the comparison vector between a set of key variables associated with $a \in A$ and $b \in B$, respectively, such as name, sex, date of birth. The key variables and the comparison vector γ_{ab} are fully observed over Ω. In cases where the key variables may be affected by errors, a match (a, b) may not have complete agreement in terms of γ_{ab}, and a non-match (a, b) can nevertheless agree on some (even all) of the key variables.

In the classical approach of Fellegi and Sunter (1969), one recognizes the probabilistic nature of γ_{ab} due to the perturbations that cause key-variable errors. The related methods are referred to as probabilistic record linkage. To explain the probabilistic record linkage method of Fellegi and Sunter (1969), let $m(\gamma_{ab}) = f(\gamma_{ab} | (a, b) \in M)$ be the probability mass function of the discrete values γ_{ab} can take given $(a, b) \in M$. Similarly, we can define $u(\gamma_{ab}) = f(\gamma_{ab} | (a, b) \in U)$. The ratio

$$r_{ab} = \frac{m(\gamma_{ab})}{u(\gamma_{ab})}$$

is then the basis of the likelihood ratio test (LRT) for $H_0: (a, b) \in M$ vs. $H_1: (a, b) \in U$. Let $M^* = \{(a, b): r_{ab} > c_M\}$ be the pairs classified as matches and $U^* = \{(a, b): r_{ab} < c_U\}$ the non-matches, the remaining pairs are classified by clerical review, where (c_M, c_U) are the thresholds related to the probabilities of false links (of pairs in U) and false non-links (of pairs in M), respectively, defined as

$$\mu = \sum_{\gamma} u(\gamma)\, \delta(M^*; \gamma) \quad \text{and} \quad \lambda = \sum_{\gamma} m(\gamma)\, \delta(U^*; \gamma), \tag{2.1}$$

where $\delta(M^*; \gamma) = 1$ if $\gamma_{ab} = \gamma$ means $(a, b) \in M^*$ and 0 otherwise, similarly for $\delta(U^*; \gamma)$.

In practice the probabilities $m(\gamma)$ and $u(\gamma)$ are unknown. Neither is the prevalence of true matches, given by $\pi = |M|/|\Omega| := n_M/n$. Let η be the set containing π and the unknown parameters of $m(\gamma)$ and $u(\gamma)$. Let $g_{ab} = 1$ if $(a,b) \in M$ and 0 if $(a,b) \in U$. Given the complete data $\{(g_{ab}, \gamma_{ab}) : (a,b) \in \Omega\}$, Winkler (1988) and Jaro (1989) assume the log-likelihood to be

$$h(\eta) = \sum_{(a,b)\in\Omega} g_{ab} \log(\pi m(\gamma_{ab})) + \sum_{(a,b)\in\Omega} (1 - g_{ab}) \log((1-\pi)\, u(\gamma_{ab})). \quad (2.2)$$

An EM-algorithm follows by treating $g_\Omega = \{g_{ab} : (a,b) \in \Omega\}$ as the missing data.

There are two fundamental problems with this classical approach.

[Problem-I] Record linkage is not a direct application of the LRT, because one needs to evaluate the pairs in Ω instead of any given pair. The classification of Ω into M^* and U^* is incoherent generally, since a given record can belong to multiple pairs. Minimum post-classification deduplication of M^* would be necessary then, which is not part of the theoretical formulation above. In particular, there lacks an associated method for estimating the uncertainty surrounding the final linked set, such as the amount of false links in it or the remaining matches outside of it.

[Problem-II] In reality the comparison vectors of any two pairs are not independent, as long as they share a record. For example, given $(a,b) \in M$ and γ_{ab} not subjected to errors, then $g_{ab'}$ must be 0, for $b' \neq b$ and $b' \in B$, as long as there are no duplicated records in either A or B, and $\gamma_{ab'}$ depends only on the key-variable errors of b'. Whereas, marginally $g_{ab'} = 1$ with probability π and $\gamma_{ab'}$ depends also on the key-variable errors of a. It follows that $h(\eta)$ in (2.2) does not correspond to the true joint data distribution of $\gamma_\Omega = \{\gamma_{ab} : (a,b) \in \Omega\}$, even when the marginal m- and u-probabilities are correctly specified. Similarly, although one may define marginally $\pi = \Pr[(a,b) \in M \,|\, (a,b) \in \Omega]$ for a randomly selected record pair from Ω, it does not follow that $\log f(g_\Omega) = n_M \log \pi + (n - n_M) \log(1-\pi)$ jointly as in (2.2). For both reasons, $h(\eta)$ given by (2.2) cannot be the complete-data log-likelihood.

In the next two sections, we develop maximum entropy classification to record linkage to avoid the problems above, after which more discussions of the classical approach will be given.

3. Maximum entropy classification: Supervised

As noted in Section 1, the record linkage problem is a classification problem. Maximum entropy classification has been used in image restoration or text analysis (Gull and Daniell, 1984; Berger, Della Pietra and Della Pietra, 1996). Maximum entropy classification (MEC) has been proposed for supervised learning (SL) to standard classification problems, where the units are known but the true

classes of the units are unknown apart from a sample of labelled units. Let $Y \in \{1, 0\}$ be the true class and $\mathbf{X}$ the random vector of features. Let the density ratio be

$$r(x;\boldsymbol{\eta}) = \frac{f(x \mid Y=1;\boldsymbol{\eta})}{f(x \mid Y=0;\boldsymbol{\eta})} := \frac{f_1(x;\boldsymbol{\eta})}{f_0(x;\boldsymbol{\eta})},$$

where f_1 and f_0 are the conditional density functions given $Y = 1$ or 0, respectively, and $\boldsymbol{\eta}$ contains the unknown parameters. For MEC based on $r(x)$, one finds $\hat{\boldsymbol{\eta}}$ that maximises the Kullback-Leibler (KL) divergence from f_0 to f_1 subjected to a constraint, i.e.

$$D = \int_{S_1} f_1(x;\boldsymbol{\eta}) \log r(\mathbf{x};\boldsymbol{\eta})\, d\mathbf{x} \quad \text{subjected to} \int_{S_1} f_0(x;\boldsymbol{\eta})\, r(\mathbf{x};\hat{\boldsymbol{\eta}})\, d\mathbf{x} = 1,$$

where S_1 is the support of X given $Y = 1$, and the normalisation constraint arises since $r(x;\boldsymbol{\eta})\, f_0(\mathbf{x};\hat{\boldsymbol{\eta}})$ is an estimate of $f_1(x)$. Provided common support $S_1 = S_0$, where S_0 is the support of X given $Y = 0$, one can use the empirical distribution function (EDF) of x over $\{x_i : y_i = 1\}$ in place of f_1 for D, and that over $\{x_i : y_i = 0\}$ in place of f_0 for the constraint. Having obtained $\hat{r}_x = r(x;\hat{\boldsymbol{\eta}})$, one can classify any unit given the associated feature vector x based on $\Pr(Y = 1 \mid x; p, \hat{r}_x)$, where p is an estimate of the prevalence $p = \Pr(Y = 1)$.

We describe how the idea of MEC for supervised learning can be adapted to record linkage problem in the following subsections.

3.1 Probability ratio for record linkage

For supervised learning based MEC to record linkage, suppose M is observed for the given Ω, and the trained classifier is to be applied to the record pairs outside Ω. To fix the idea, suppose B is a non-probability sample that overlaps with the population P, and A is a probability sample from P with known inclusion probabilities. While $\boldsymbol{\gamma}_M = \{\gamma_{ab} : (a,b) \in M\}$ may be considered as an IID sample, since each (a,b) in M refers to a distinct entity, this is not the case with $\{\gamma_{ab} : (a,b) \notin M\}$, whose joint distribution is troublesome to model.

Probability ratio (I)

Let $r_q(\gamma)$ be the probability ratio given by

$$r_q(\gamma) = \frac{m(\gamma)}{q(\gamma)},$$

where $m(\gamma)$ is the probability mass function of $\gamma_{ab} = \gamma$ given $g_{ab} = 1$, and $q(\gamma)$ is that over $\boldsymbol{\gamma}_\Omega = \{\gamma_{ab} : (a,b) \in \Omega\}$. The KL divergence measure from $q(\gamma)$ to $m(\gamma)$ and the normalisation constraint are

$$D_f = \sum_{\gamma \in S(M)} m(\gamma) \log r_q(\gamma) \quad \text{and} \quad \sum_{\gamma \in S(M)} q(\gamma) = 1,$$

where $S(M)$ is the support of γ_{ab} given $g_{ab} = 1$. This setup allows $S(M)$ to be a subset of S, where S is the support of all possible γ_{ab}. It follows that, based on the IID sample γ_M of size $n_M = |M|$, the objective function to be minimized for r_q can be given by

$$Q_f = \sum_{(a,b) \in M} \frac{f(\gamma_{ab})}{n_M(\gamma_{ab})} r_q(\gamma_{ab}) - \frac{1}{n_M} \sum_{(a,b) \in M} \log r_q(\gamma_{ab}), \tag{3.1}$$

where $f(\gamma_{ab}) = \sum_{(i,j) \in M} t_k(\gamma_{ij})$ and $n_M(\gamma_{ab})$ based on the observed support $S(M)$.

Probability ratio (II)

Provided $S(M) \subseteq S(U)$, where $S(U)$ is the support of γ_{ab} over U, one can let the probability ratio be given by

$$r(\gamma) = \frac{m(\gamma)}{u(\gamma)}$$

where $u(\gamma)$ is the probability of $\gamma_{ab} = \gamma$ given $g_{ab} = 0$. We have

$$r_q(\gamma) = \frac{m(\gamma)}{q(\gamma)} = \frac{m(\gamma)}{\pi m(\gamma) + (1-\pi) u(\gamma)} = \frac{r(\gamma)}{\pi(r(\gamma) - 1) + 1}$$

where $q(\gamma) = \pi m(\gamma) + (1-\pi) u(\gamma)$, so that $r_q(\gamma)$ and $r(\gamma)$ are one-to-one. Meanwhile, the KL divergence measure from $u(\gamma)$ to $m(\gamma)$ is given by

$$D = \sum_{\gamma \in S(M)} m(\gamma) \log r(\gamma)$$

and the objective function to be minimized for r can now be given by

$$Q = \sum_{(a,b) \in M} \frac{u(\gamma_{ab})}{n_M(\gamma_{ab})} r(\gamma_{ab}) - \frac{1}{n_M} \sum_{(a,b) \in M} \log r(\gamma_{ab}). \tag{3.2}$$

Model of γ: Under the multinomial model, one can simply use the EDF of over γ_Ω as $f(\gamma)$, for each distinct level of γ, as long as $|\Omega|$ is large compared to $|S|$. Similarly for $m(\gamma)$ over γ_M and $u(\gamma)$ over U. For linkage outside of Ω, the estimated $m(\gamma)$ from $M(\Omega)$ applies, if the selection of A from P is non-informative.

For γ made up of K binary agreement indicators $\gamma_k = 0, 1$ for $k = 1, \ldots, K$, there are up to 2^K distinct levels of γ, which can sometimes be relatively large compared to $|M|$. A more parsimonious model of $m(\gamma; \theta)$ that is commonly used is given by

$$m(\gamma; \theta) = \prod_{k=1}^{K} \theta_k^{\gamma_k} (1 - \theta_k)^{1-\gamma_k} \tag{3.3}$$

where $\theta_k = \Pr(\gamma_{ab,k} = 1 \mid g_{ab} = 1)$, and $\gamma_{ab,k}$ is the k^{th} component of γ_{ab}. It is possible to model θ_k based on the distributions of the key variables that give rise to γ_k, which makes use of the differential frequencies of their values, such as the fact that some names are more common than others. Similarly, $u(\gamma;\xi)$ can be modeled as in (3.3) with parameters ξ_k instead of θ_k, where $\xi_k = \Pr(\gamma_{ab,k} = 1 \mid g_{ab} = 0)$.

Note that (3.3) implies conditional independence among agreement indicators. Winkler (1993) and Winkler (1994) demonstrated that even when the conditional independence assumption does not hold, results based on conditional independence assumption are quite robust. More complicated models that allow for correlated γ_k can also be considered. See Armstrong and Mayda (1993) and Larsen and Rubin (2001) for discussion of those models. See Xu, Li, Shen, Hui and Grannis (2019) for a recent study which compares models with or without correlated γ_k.

3.2 MEC sets for record linkage

Provided there are no duplicated records in either A or B, a classification set for record linkage, denoted by M, consists of record pairs from Ω, where any record in A or B appears at most in one record pair in M. Let the entropy of a classification set M be given by

$$D_M = \frac{1}{|M|} \sum_{(a,b) \in M} \log r(\gamma_{ab}). \tag{3.4}$$

A MEC set of given size $n^* = |M|$ is the first classification set that is of size n^*, obtained by deduplication in the descending order of $r(\gamma_{ab})$ over Ω. It is possible to have $(a,b') \notin M$ and $r(\gamma_{ab'}) > r(\gamma_{a',b'})$ for $(a',b') \in M$, if there exists $(a,b) \in M$ with $r(\gamma_{ab}) > r(\gamma_{ab'})$.

A MEC set of size n^* is not necessarily the largest possible classification set with the maximum entropy, to be referred to as a maximal MEC set, which is the largest classification set such that $r(\gamma_{ab}) = \max_\gamma r(\gamma)$ for every (a,b) in it. In practice, a maximal MEC set is given by the first pass of deterministic linkage, which only consists of the record pairs with perfect and unique agreement of all the key variables.

Probabilistic linkage methods for MEC set are useful if one would like to allow for additional links, even though their key variables do not agree perfectly with each other. For the uncertainty associated with a given MEC set M, we consider two types of errors. First, we define the false link rate (FLR) among the links in M to be

$$\psi = \frac{1}{|M|} \sum_{(a,b) \in M} (1 - g_{ab}) \tag{3.5}$$

which is different to μ by (2.1) where the denominator is $|U|$. Second, the missing match rate (MMR) of M, which is related to the false non-link probability λ in (2.1), is given by

$$\tau = 1 - \frac{1}{n_M} \sum_{(a,b) \in M} g_{ab}. \tag{3.6}$$

While μ and λ in (2.1) are theoretical probabilities, the FLR and MMR are actual errors.

It is instructive to consider the situation, where one is asked to form MEC sets Ω given all the necessary estimates related to the probability ratio $r(\gamma)$, which can be obtained under the SL setting, without being given n_M, g_Ω or M directly.

First, the perfect MEC set should have the size n_M. Let $J := \sum_{(a,b)\in\Omega} t := \sum_{(a,b)\in\Omega} L$. One can obtain n_M as the solution to the following fixed-point equation:

$$n_M = \sum_{(a,b)\in\Omega} g(\gamma_{ab}) = \sum_{\gamma\in S} n(\gamma)\, \hat{g}(\gamma) \tag{3.7}$$

where

$$\hat{g}(\gamma) := \Pr\left(g_{ab}=1 \mid \gamma_{ab}=\gamma\right) = \frac{\pi r(\gamma)}{\pi\left(r(\gamma)-1\right)+1} = \frac{n_M r(\gamma)}{n_M\left(r(\gamma)-1\right)+n} \tag{3.8}$$

and the probability is defined with respect to completely random sampling of a single record pair from Ω. To see that $\hat{g}(\gamma)$ by (3.8) satisfies (3.7), notice $\hat{g}(\gamma) = n_M m(\gamma)/n_M$ satisfies (3.7) for any well-defined $m(\gamma)$, and $n(\gamma)/n = \pi m(\gamma) + (1-\pi)u(\gamma)$ by definition.

Next, apart from a maximal MEC set, one would need to accept discordant pairs. In the SL setting, one observes the EDF of γ over M, giving rise to $\hat{\theta}_k = n_M(1;k)/n_M$, where $n_M(1;k)$ is the number of agreements on the k^{th} key variable over M. The perfect MEC set M^* should have these agreement rates. We have then, for $k = 1,\ldots,K$,

$$\theta_k^* = \frac{1}{|M^*|}\sum_{(a,b)\in M^*} t_k := \frac{1}{J}\sum_{(a,b)\in\Omega} t_k \quad \text{so for} \quad |M^*| = n_M. \tag{3.9}$$

Thus, no matter how one models $r(\gamma)$, the perfect MEC set should satisfy jointly the $K+1$ equations defined by (3.7) and (3.9), given the knowledge of $r(\gamma)$.

4. MEC for unsupervised record linkage

Let z be the K-vector of key variables, which may be imperfect for two reasons: it is not rich enough if the true z-values are not unique for each distinct entity underlying the two files to be linked, or it may be subjected to errors if the observed z is not equal to its true value. Let A contain only the distinct z-vectors from the first file, after removing any other record that has a duplicated z-vector to some record that is retained in A. In other words, if the first file initially contains two or more records with exactly the same value of the combined key, then only one of them will be retained in A for record linkage to the second file. Similarly let B contain the unique records from the second file. The reason for separate deduplication of keys is that no comparisons between the two files can distinguish among the duplicated z in either file, which is an issue to be resolved otherwise.

Given A and B preprocessed as above, the maximal MEC set M_1 only consists of the record pairs with the perfect agreement of all the key variables. For probabilistic linkage beyond M_1, one can follow the same scheme of MEC in the supervised setting, as long as one is able to obtain an estimate of probability ratio, given which one can form the MEC set of any chosen size. Nevertheless, to estimate the associated FLR (3.5) and MMR (3.6), an estimate of n_M is also needed.

4.1 Algorithm of unsupervised MEC

The idea now is to apply (3.7) and (3.9) jointly. Since setting $n_M = |M_1|$ and $\theta_k = 1$ associated with the maximal MEC set satisfies (3.7) and (3.9) automatically, probabilistic linkage requires one to assume $n_M > |M_1|$ and $\theta_k < 1$ for at least some of $k = 1, \ldots, K$. Moreover, unless there is external information that dictates it otherwise, one can only assume common support $S(M) = S(U)$ in the unsupervised setting. Let

$$r(\gamma) = m(\gamma; \boldsymbol{\theta}) / u(\gamma; \xi) \tag{4.1}$$

where the probability of observing γ is $m(\gamma; \boldsymbol{\theta})$ by (3.3) given that a randomly selected record pair from Ω belongs to M, and $u(\gamma; \xi)$ otherwise, similarly given by (3.3) with parameters ξ_k instead of θ_k. An iterative algorithm of unsupervised MEC is given below.

I. Set $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_K^{(0)})$ and $n_M^{(0)} = |M_1|$, where M_1 is the maximal MEC set.

II. For the t^{th} iteration, let $g_{ab}^{(t)} = 1$ if $(a, b) \in M^{(t)}$, and 0 otherwise.

i. Update $u(\gamma; \xi^{(t)})$ by using (4.4), which is discussed below, given $g^{(t)} = \{g_{ab}^{(t)} : (a, b) \in \Omega\}$, and calculate

$$\theta_k^{(t)} = \frac{1}{|M^{(t)}|} \sum_{(a,b) \in \Omega} g_{ab}^{(t)} \quad \text{for } k = 1, \ldots, K \tag{4.2}$$

which maximize D_M in (3.4) for given $u(\gamma; \xi^{(t)})$, $M^{(t)} = \{(a, b) \in \Omega : g_{ab}^{(t)} = 1\}$ and $|M^{(t)}| = \sum_{(a,b) \in \Omega} g_{ab}^{(t)}$. Once $\boldsymbol{\theta}^{(t)}$ and $\xi^{(t)}$ are obtained, we can update $n_M^{(t)} = \sum_\gamma n(\gamma)\, g^{(t)}(\gamma)$, where

$$g^{(t)}(\gamma) \equiv g(\gamma; \boldsymbol{\theta}^{(t)}, \xi^{(t)}) = \min\left\{ \frac{|M^{(t)}| r^{(t)}(\gamma)}{|M^{(t)}| (r^{(t)}(\gamma) - 1) + n}, 1 \right\}$$

$$r^{(t)}(\gamma) \equiv r(\gamma; \boldsymbol{\theta}^{(t)}, \xi^{(t)}) = \frac{m(\gamma; \boldsymbol{\theta}^{(t)})}{u(\gamma; \xi^{(t)})}.$$

ii. For given $\boldsymbol{\theta}^{(t)}, \xi^{(t)}$ and $n_M^{(t)}$, we find the MEC set $M^{(t+1)} = \{(a, b) \in \Omega : g_{ab}^{(t+1)} = 1\}$ such that $|M^{(t+1)}| = n_M^{(t)}$ by deduplication in the descending order of $r^{(t)}(\gamma_{ab})$ over Ω. It maximizes the entropy denoted by $Q^{(t)}(g)$:

$$Q^{(t)}(g) \equiv Q\left(g \mid \psi^{(t)}\right) = \frac{1}{n_M^{(t)}} \sum_{(a,b)\in\Omega} g_{ab} \log r^{(t)}(\gamma_{ab}), \tag{4.3}$$

with respect to g.

III. Iterate until $n_M^{(t)} = n_M^{(t+1)}$ or $\left\| \theta^{(t)} - \theta^{(t+1)} \right\| < \delta$, where δ is a small positive value.

A theoretical convergence property of the proposed algorithm and its proof are presented in the supplementary materials.

Notice that, insofar as $\Omega = M \cup U$ is highly imbalanced, where the prevalence $g_{ab} = 1$ is very close to 0, one could simply ignore the contributions from M and use

$$\xi_k = \frac{1}{n} \sum_{(a,b)\in\Omega} t_k(\gamma_{ab}), \text{ so} \tag{4.4}$$

under the model (3.3) of $u(\gamma;\xi)$, in which case there is no updating of $u(\gamma;\xi^{(t)})$. Other possibilities of estimating $u(\gamma;\xi)$ will be discussed in Section 5.2.

Table 4.1 provides an overview of MEC for record linkage in the supervised or unsupervised setting. In the supervised setting, one observes γ for the matched record pairs in M, so that the probability $m(\gamma)$ can be estimated from them directly. Whereas, for MEC in the unsupervised setting, one cannot separate the estimation of $m(\gamma)$ and n_M.

Table 4.1
MEC for record linkage in supervised or unsupervised setting

	Supervised	Unsupervised
$\Omega = M \cup U$	Observed	Unobserved
Probability ratio	$r_q(\gamma)$ generally applicable $r(\gamma)$ given $S(M) \subseteq S(U)$	$r(\gamma)$ generally assuming $S(M) = S(U)$
Model of γ	Multinomial if only discrete comparison scores Directly or via key variables and measurement errors	
MEC set	Guided by FLR and MMR Require estimate of n_M in addition	
Estimation	$m(\gamma;\theta)$ from γ_M in Ω n_M by (3.7) outside Ω	$m(\gamma;\theta)$ and n_M jointly by (3.7) and (3.9)

4.2 Error rates

MEC for record linkage should generally be guided by the error rates, FLR and MMR, without being restricted to the estimate of n_M.

Note that $\{g_{ab}: (a,b) \in M\}$ of any MEC set M are among the largest ones over Ω, because MEC follows the descending order of g_{ab}, except for necessary deduplication when there are multiple pairs involving a given record. To exercise greater control of the FLR, we let the target FLR, and consider the following bisection procedure.

i. Choose a threshold value c_ψ and form the corresponding MEC set $M(c_\psi)$, where $r_{ab} \geq c_\psi$ for any $(a, b) \in M(c_\psi)$.

ii. Calculate the estimated FLR of the resulting MEC set as

$$\hat{\psi} = \frac{1}{|M|} \sum_{(a,b) \in M} (1 - q_{ab}). \tag{4.5}$$

If $\hat{\psi} > \psi$, then increase c_ψ; if $\hat{\psi} < \psi$, then reduce c_ψ.

Iteration between the two steps would eventually lead to a value of c_ψ that makes $\hat{\psi}$ as close as possible to ψ, for the given probability ratio $r(\gamma)$.

The final MEC set $\hat{M}$ can be chosen in light of the corresponding FLR estimate. It is also possible to take into consideration the estimated MMR given by

$$\hat{\tau} = 1 - \sum_{(a,b) \in M} q_{ab} / n_M \tag{4.6}$$

where n_M is given by unsupervised MEC algorithm. Note that if $|M| = n_M$, then we shall have $\hat{\psi} = \hat{\tau}$, but not if $\hat{M}$ is guided by a given target value of FLR or MMR.

In Section 6.2, we investigate the performance of the MEC sets guided by the error rates through simulations.

5. Discussion

Below we discuss and compare two other approaches in the unsupervised setting, including the ways by which some of their elements can be incorporated into the MEC approach. Other less practical approaches are discussed in the supplementary material.

5.1 The classical approach

Recall Problems I and II of the classical approach mentioned in Section 2.

From a practical point of view, Problem I can be dealt with by any deduplication method of the set M^* of classified records pairs, where $r(\gamma_{ab})$ is above a threshold value for all $(a, b) \in M^*$. \$V ³DQ DGYDQFH RYHU SUHYLRXV DG KRF Jaro (1989) chooses the linked set $M \subseteq M^*$, which maximises the sum of $\log r(\gamma_{ab})$ subject to the constraint of one-to-one link. Since q_{ab} is a monotonic function of $r(\gamma_{ab})$, this amounts to choosing M which maximises the expected number of matches in it, denoted by

$$n_M^* = \sum_{(a,b) \in M} q_{ab}$$

But n_M^* is still not connected to the probabilities of false links and non-links defined by (2.1). As illustrated below, neither does it directly control the errors of the linked set M.

&RQVLGHU OLQNLQJ WZR ILOHV ZLWK UHFRUGV HDFK 6X $ $ $

perturbation that could potentially cause disagreement of the k^{th} key variable, and $e_{ab,k} = 0$ otherwise. Let

$$\theta_k = (1-\alpha_k) + \alpha_k \sum_{d=1}^{D_k} m_{kd}^2 = 1 - \alpha_k \left(1 - \sum_{d=1}^{D_k} m_{kd}^2 \right)$$

where we assume that α_k must be positive for some $k = 1, \ldots, K$, and

$$m_{kd} = \Pr(z_{ik} = d \mid g_{ab} = 1, e_{ab,k} = 1) = \Pr(z_{ik} = d \mid g_{ab} = 1, e_{ab,k} = 0)$$

for $i = a$ or b. Next, for any record in either A or B, let $\delta_i = 1$ if it has a match in the other file and $\delta_i = 0$ otherwise. Given $\delta_i = 0$, with or without perturbation, let $\Pr(z_{ik} = d \mid \delta_i = 0) = u_{kd}$. We have $\beta_{kd} := m_{kd} \equiv u_{kd}$ if δ_i is non-informative. A slightly more relaxed assumption is that it is only non-informative in one of the two files. To be more resilient against its potential failure, one can assume u_{kd} to hold for all the records in the smaller file, and allow u_{kd} to differ for the records with $\delta_i = 0$ in the larger file. Suppose $n_A < n_B$. Let

$$p = \Pr(\delta_b = 1) = E(n_M) / n_B = n_A \pi$$

be the probability that a record in B has a match in A. One may assume $z_A = \{z_a : a \in A\}$ to be independent over A, giving

$$\ell_A = \sum_{a \in A} \sum_{k=1}^{K} \log m_{ak}$$

where $m_{ak} = \sum_{d=1}^{D_k} m_{kd}\, I(z_{ak} = d)$. The complete-data log-likelihood based on (δ_B, z_B) is

$$\ell_B = \sum_{b \in B} \delta_b \log \left(p \prod_{k=1}^{K} m_{bk} \right) + \sum_{b \in B} (1-\delta_b) \log \left((1-p) \prod_{k=1}^{K} u_{bk} \right) \tag{5.2}$$

where $m_{bk} = \sum_{d=1}^{D_k} m_{kd}\, I(z_{bk} = d)$ and $u_{bk} = \sum_{d=1}^{D_k} u_{kd}\, I(z_{bk} = d)$, based on an assumption of independent (δ_b, z_b) across the entities in B.

Under separate modelling of z_A and (z_B, δ_B), let $\hat{m}_{kd}$ be the MLE based on z_A, given which an EM-algorithm for estimating p and u_{kd} follows from (5.2) by treating δ_B as the missing data. However, the estimation is feasible only if $\{u_{kd}\}$ and $\{m_{kd}\}$ are not exactly the same; whereas the MLE $\hat{n}_M$ of n_M has a large variance, when $\{m_{kd}\}$ and $\{u_{kd}\}$ are close to each other, even if they are not exactly equal.

Meanwhile, the closeness between $\{m_{kd}\}$ and $\{u_{kd}\}$ does not affect the MEC approach, where $\hat{n}_M$ is obtained from solving (3.7) given $r(\hat{\gamma}) = m(\hat{\gamma}) / u(\gamma)$, where $u(\hat{\gamma})$ is indeed most reliably estimated when $\{m_{kd}\} = \{u_{kd}\}$. Moreover, one can incorporate a profile EM-algorithm, based on (5.2) given $n_M^{(t)}$, to update $u(\gamma; \xi^{(t)})$ in the unsupervised MEC algorithm of Section 4.1. At the t^{th} iteration, where $t \geq 1$, given $p^{(t)} = n_M^{(t)} / \max(n_A, n_B)$ and $\hat{m}_{kd}$ estimated from the smaller file A, obtain $u_{kd}^{(t)}$ by

$$\xi_k^{(t)} = \left((1-p^{(t)}) \sum_{d=1}^{D_k} u_{kd}^{(t)} m_{kd} + p^{(t)} \left(1 - \frac{1}{n_A} \right) \sum_{d=1}^{D_k} \hat{m}_{kd}^2 \right) \Big/ \left(1 - p^{(t)} / n_A \right). \tag{5.3}$$

6. Simulation study

6.1 Set-up

To explore the practical feasibility of the unsupervised MEC algorithm for record linkage, we conduct a simulation study based on the data sets listed in Table 6.1, which are disseminated by ESSnet DI (McLeod, Heasman and Forbes, 2011) and freely available online. Each record in a data set has associated synthetic key variables, which may be distorted by missing values and typos when they are created, in ways that imitate real-life errors (McLeod et al., 2011).

Table 6.1
Data set description (size in parentheses)

Data set	Description
Census (25,343)	A fictional data set to represent some observations from a decennial Census
CIS (24,613)	Fictional observations from Customer Information System, combined administrative data from the t benefit systems
PRD (24,750)	Fictional observations from Patient Register Data of the National Health Service

We consider the linkage keys forename, surname, sex, and date of birth (DOB). To model the key variables, we divide DOB into 3 key variables (Day, Month, Year). For text variables such as forename and surname, we divide them into 4 key variables by using the Soundex coding algorithm (Gotlieb and Hilton, 1990, page 290), which reduces a name to a code consisting of the leading letter followed by three digits, e.g. Copas ≡ C120, Hilton ≡ H435. The twelve key variables for record linkage are presented in Table 6.2.

Table 6.2
Twelve key variables available in the three data sets

Variable		Description	No. of Categories
PERNAME1	1	First letter of forename	26
	2	First digit of Soundex code of forename	7
	3	Second digit of Soundex code of forename	7
	4	Third digit of Soundex code of forename	7
PERNAME2	1	First letter of surname	26
	2	First digit of Soundex code of surname	7
	3	Second digit of Soundex code of surname	7
	4	Third digit of Soundex code of surname	7
SEX		Male/Female	2
DOB	DAY	Day of birth	31
	MON	Month of birth	12
	YEAR	Year of birth (1910 ~ 2012)	103

We set up two scenarios to generate linkage files. We use the unique identification variable (PERSON ID) for sampling, which are available in all the three data sets. We sample $n_A = 5{,}000$ and $n_B = 1{,}000$ individuals from PRD and CIS, respectively. Let p_A be the proportion of records in the smaller file (PRD) that are also selected in the larger file (CIS), by which we can vary the degree of overlap, i.e. the set of matched individuals AB, between A and B. We use $p_A = 0.8$, 0.5 or 0.3 under either scenario.

Scenario I (Non-informative)

- Sample $n_0 = n_B / p_A$ individuals randomly from Census.
- Sample n_A randomly from these n_0 as the individuals of PRD, denoted by A.
- Sample n_B randomly from these n_0 as the individuals of CIS, denoted by B.

Under this scenario both δ_a and δ_b are non-informative for the key-variable distribution. For any given p_A, we have $E(n_M) = n_A p_A$ and $\pi = E(n_M) / n_0$, where n_M is the random number of matched individuals between the simulated files A and B.

Scenario II (Informative)

- Sample n_A randomly from Census $\cap$ PRD $\cap$ CIS, denoted by A from PRD.
- Sample $n_M = n_A p_A$ randomly from A as the matched individuals, denoted AB.
- Sample $n_B - n_M$ randomly from CIS $\setminus$ A having SEX = F, YEAR $\leq$ 1970, and odd MON, denoted by B_0. Let $B = AB \cup B_0$ be the sampled individuals of CIS.

Under this scenario the key-variable distribution is the same in A, whether or not $\delta_a = 1$, but it is different for the records $b \in B_0$, or $\delta_b = 0$. Hence, scenario II is informative. For any given p_A, we have fixed $n_M = n_A p_A$ and $\pi = p_A / n_B$.

6.2 Results: Estimation

For the unsupervised MEC algorithm given in Section 4.1, one can adopt (4.2) or (5.1) for updating $\theta_k^{(t)}$. Moreover, one can use (4.4) for $\xi_{ik}^{(t)}$ directly, or (5.3) for updating $\xi_{ik}^{(t)}$ iteratively. In particular, choosing (5.1) and (4.4) effectively incorporates the procedure of Winkler (1988) and Jaro (1989) for parameter estimation. Note that the MEC approach still differs to that of Jaro (1989) with respect to the formation of the linked set M.

Table 6.3 compares the performance of the unsupervised MEC algorithm, using different formulae for $\theta_k^{(t)}$ and $\xi_{ik}^{(t)}$, where the size of $M^{(t)}$ is equal to the corresponding estimate of n_M. In addition, we include $\theta_k^{(0)} = n_M(1; k) / n_M$ estimated directly from the matched pairs in M, as if M were available for supervised learning, together with (4.4) for $\xi_{ik}^{(0)}$. The true parameters and error rates are given in addition to their estimates.

Table 6.3
Parameters and averages of their estimates, averages of error rates and their estimates, over 200 simulations
Median of estimate of n_M given as $\tilde{n}_M$

Scenario I										Scenario II									
Parameter		Formulae		Estimation						Parameter		Formulae		Estimation					
π	$E(n_M)$	$\theta_k^{(t)}$	$\xi_k^{(t)}$	$\hat{\pi}$	$\hat{n}_M$	$\tilde{n}_M$	FLR	MMR	$\widehat{\text{FLR}}$	$\widehat{\text{MMR}}$	π	n_M	$\theta_k^{(t)}$	$\xi_k^{(t)}$	$\hat{\pi}$	$\hat{n}_M$	$\tilde{n}_M$	FLR	MMR	$\widehat{\text{FLR}}$	$\widehat{\text{MMR}}$
0.0008	400	$\hat{\theta}_k$	(4.4)	0.00080	400.0	397	0.0264	0.0266	0.0357	0.0357	0.0008	400	$\hat{\theta}_k$	(4.4)	0.00080	398.3	400	0.0230	0.0273	0.0326	0.0326
		(4.2)	(5.3)	0.00082	407.9	405	0.0425	0.0257	0.0509	0.0509			(4.2)	(5.3)	0.00080	401.4	401	0.0305	0.0277	0.0403	0.0403
		(4.2)	(4.4)	0.00083	414.7	407	0.0549	0.0244	0.0620	0.0620			(4.2)	(4.4)	0.00081	405.2	404	0.0379	0.0262	0.0467	0.0467
		(5.1)	(4.4)	0.00081	406.0	405	0.0399	0.0269	0.0503	0.0503			(5.1)	(4.4)	0.00080	401.4	401	0.0316	0.0286	0.0438	0.0438
0.0005	250	$\hat{\theta}_k$	(4.4)	0.00050	251.6	249	0.0340	0.0301	0.0370	0.0370	0.0005	250	$\hat{\theta}_k$	(4.4)	0.00050	249.6	250	0.0284	0.0302	0.0334	0.0334
		(4.2)	(5.3)	0.00052	258.3	255	0.0559	0.0296	0.0533	0.0533			(4.2)	(5.3)	0.00050	251.8	251	0.0383	0.0320	0.0410	0.0410
		(4.2)	(4.4)	0.00053	266.9	256.5	0.0742	0.0277	0.0680	0.0680			(4.2)	(4.4)	0.00052	257.7	253	0.0513	0.0295	0.0516	0.0516
		(5.1)	(4.4)	0.00052	261.7	259	0.0676	0.0305	0.0636	0.0636			(5.1)	(4.4)	0.00051	255.4	253.5	0.0510	0.0336	0.0520	0.0520
0.0003	150	$\hat{\theta}_k$	(4.4)	0.00030	152.3	151	0.0439	0.0356	0.0381	0.0381	0.0003	150	$\hat{\theta}_k$	(4.4)	0.00030	150.5	150	0.0382	0.0355	0.0350	0.0350
		(4.2)	(5.3)	0.00033	165.9	156.5	0.0873	0.0244	0.0620	0.0620			(4.2)	(5.3)	0.00031	153.0	153	0.0559	0.0377	0.0452	0.0452
		(4.2)	(4.4)	0.00041	205.4	161	0.1632	0.0308	0.1251	0.1251			(4.2)	(4.4)	0.00032	158.5	155	0.0708	0.0342	0.0558	0.0558
		(5.1)	(4.4)	0.00054	271.4	169	0.3015	0.0785	0.1639	0.1639			(5.1)	(4.4)	0.00038	189.3	156	0.1414	0.0524	0.0903	0.0903

As expected, the best results are obtained when the parameter θ_k is estimated directly from the matched pairs in M, i.e., $\hat{\theta}_k = n_M(1;k)/n_M$, together with (4.4) for ξ_k, despite $\xi_k^{(t)}$ by (4.4) is not exactly unbiased. Nevertheless, the approximate estimator can be improved, since the profile EM estimator given by (5.3) is seen to perform better across all the tests, where both are combined with (4.2) for $\theta_k^{(t)}$. When it comes to the two formulae of $\theta_k^{(t)}$ by (4.2) and (5.1), and the resulting n_M-estimators and the error rates FLR and MMR, we notice the followings.

- Scenario I: When the size of the matched set M is relatively large at $p_A = 0.8$, there are only small differences in terms of the average and median of the two estimators of n_M, and the difference is just a couple of false links in terms of the linkage errors. Figure 6.1 shows that (4.2) results in a few larger errors of $\hat{n}_M$ than (5.1) over the 200 simulations, when $p_A = 0.8$ or $\pi = 0.0008$. As the size of the matched set M decreases, the averages and medians of the estimators of n_M resulting from (4.2) and (5.3) are closer to the true values than those of the other estimators. Especially when the matched set M is relatively small, where $\pi = 0.0003$, the formula (5.1) results in considerably worse estimation of n_M in every respect. While this is partly due to the use of (4.4) instead of (5.3), most of the difference is down to the choice of $\theta_k^{(t)}$, which can be seen from intermediary comparisons to the results based on (4.2) and (4.4).
- Scenario II: The use of (4.2) and (5.3) for the unsupervised MEC algorithm performs better than using the other formulae in terms of both estimation of n_M and error rates across the three sizes of the matched set (Figure 6.2). Relatively greater improvement is achieved by using (4.2) and (5.3) for the smaller matched sets.

The results suggest that the unsupervised MEC algorithm tends to be more affected by the size of the matched set under Scenario I than Scenario II. Choosing (4.2) and (5.3), however, seems to yield the most