Overview of Data Linkage Methods for Policy Design and Evaluation

Natalie Shlomo

1 Introduction

Data and technology are the building blocks of evidence-based policy. With the increasing availability of government service administrative data and the launch of dedicated research centres such as the United Kingdom (UK) Administrative Data Research Network, the potential for using data and technology to inform policy issues has never been greater.

A key technological tool to exploit the wealth of information contained in administrative data and other data sources is data linkage. In its simplest form, data linkage brings together information from two different records that are believed to belong to the same entity based on a set of identifiers or quasi-identifiers, known as matching variables. As mentioned, the distinction made here is that the aim is to link records for the same entity. This distinguishes the approach taken here from other recently developed methods that have been used to integrate data sources, known as data fusion or statistical matching (D'Orazio et al. 2006).

This chapter focuses briefly on deterministic (exact) matching, where all variables have to match exactly to determine a match. It then focuses on probabilistic data linkage, where allowances are made for errors in the matching variables. There are three possible scenarios in this type of linkage:

- If two records agree on all matching variables, it is unlikely that they would have agreed by chance the level of assurance that the link is correct will be high, and it is assumed that the record pair belongs to the same entity.
- If all of the matching variables disagree, the pair will not be linked as a match, and it is unlikely that the record pair belongs to the same entity.

N. Shlomo (✉)
Social Statistics Department, University of Manchester, Manchester, UK
e-mail: natalie.shlomo@manchester.ac.uk

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_4

- If there are intermediate situations where some matching variables agree and some matching variables disagree, it is necessary to predict whether the pair is a true match or a non-match. Often clerical intervention will be needed to determine the match status.

The challenge in data linkage is when there are errors in matching variables and no unique high-quality identifier such as an ID number is available. In that case, use is made of probabilistic data linkage with matching variables such as name, year of birth or place of residence, which may be prone to error. When combined and concatenated, matching variables should identify an entity uniquely across the data sources (unless the aim is to deduplicate one file). They also need to be accurate and stable over time, so place of residence can be problematic if the timeliness of the two files to be matched is not considered. In addition, there may be differences in how the data is captured and maintained in different databases.

Therefore, the key technical challenges when carrying out a probabilistic data linkage application are the following:

- The availability of good-quality identifiers to discriminate between the entity to whom the record refers and all other entities
- Deciding whether or not discrepancies in identifiers are due to mistakes in reporting for a single entity
- Processing a large volume of data within a reasonable amount of computer processing time

Sections 2 and 3 focus on deterministic (exact) matching and probabilistic record linkage, explained through the three stages of linkage: pre-linkage, linkage and post-linkage. Section 4 describes some recent advances in research related to data linkage, and Sect. 5 provides an overview of methods for the analysis of linked data that may be subject to linkage errors.

2 Deterministic (Exact) Matching Method

In deterministic (exact) matching, the records in two datasets must agree exactly on every character of every matching variable to conclude that they correspond to the same entity. It is generally used when a high-quality identifier such as an ID number is available. If there is no ID number, matching variables, such as age, gender, place of residence, etc. can be concatenated to form a unique ID number, but these variables may be prone to error.

Deterministic matching assumes that there is no error in the way the data is recorded and captured. Information may be missing or inaccurate, and there may be variations in format or inaccuracies in spelling across different sources. For this reason, relaxations have been proposed allowing some errors to be taken into account. For example, first and last names can be transformed into a phonetic code, or the names can be truncated, e.g. by using the first five letters of a name only.

Decision rules can also be set where, if there is an agreement on most of the matching variables, the pair will be declared a match. In deterministic matching, all matching variables have equal weights associated with them, such that an agreement on gender would have the same contribution to the overall decision on a correct match as an agreement on last name, although the latter should clearly contribute more to the decision.

Another important feature is that deterministic matching carries out a one-to-one match, and there are only two possible outcomes of the decision: match or no match. It is useful to carry out a deterministic matching procedure before moving to the probabilistic data linkage if the aim is to carry out a one-to-one match on a set of matching variables, since this may reduce the overall computational burden. Manual review is still needed following deterministic matching to carry out checks for any linkage errors.

3 Probabilistic Data Linkage

A probabilistic data linkage application typically involves three stages:

- Pre-linkage: editing and data cleaning, parsing fused strings such as first and last name or house number and street name, and standardising matching variables so that they have the same formats and definitions.
- Linkage: bringing pairs together for comparison and determining correct matches, i.e. the pair belongs to the same entity.
- Post-linkage: checking residuals for the unmatched, determining error rates and other quality indicators and carrying out analysis taking into account linkage errors.

The following sections describe the probabilistic data linkage method in terms of these three stages: pre-linkage, linkage and post-linkage.

3.1 Pre-linkage Stage

3.1.1 Data Standardisation

The success of data linkage depends on the quality of the data. Pre-processing and data cleaning are the most difficult and time-consuming steps in data linkage, but they are necessary to ensure a successful and accurate linkage. In the first step, a reference number needs to be generated and added to each record across the files to be linked. The reference number should contain a header denoting the iteration of the linkage. Duplicates need to be removed (unless the aim of the linkage is to deduplicate a dataset).

Matching variables need to be selected. The choice depends on the type and contents of the datasets. When no stable ID number is available, it is necessary to link on less stable variables, which may be subject to errors and omissions. The criteria for matching variables are uniqueness, availability, accuracy and stability over time.

Some considerations for matching variables are:

- Proper names rarely change during the lifetime of a person, for example, birth surname, first forename and initials.
- Personal characteristics that are fixed at birth very rarely change, for example, gender, ethnicity, date of birth, place of birth and social security number.
- Social demographic variables may change over time, for example, street name, postcode, marital status, social class and date of marriage.

Matching variables should have high discriminating power. Examples of variables with high discriminating power are those with a large number of value states, such as zip code and last name. Examples of variables with low discriminating power are those with a small number of value states, such as gender and month of birth.

All datasets should be checked for completeness with a clear understanding of the coverage of each of the datasets. Variables involved in the data linkage should be free of errors. Some errors can be detected by checking logical consistencies in the data; for example, a marital status of 'married' is not possible for an individual under the age of 14. Ranges of numerical variables and check digits of ID numbers can be easily verified. Other edits and corrections are:

- Matching variables might include fused strings, such as first name given together with last name, or house number given together with street name. These matching variables need to be parsed into separate matching variables, as this increases the power of the decision rule to determine correct matches.
- Names may suffer from variations in spelling, or the use of nicknames and abbreviations, and this increases the complexity of the linkage. This is typically solved by using dictionaries that equate different versions of names and can be tailored to different cultures that use different nicknames, for example, William and Bill. In addition, spelling variations of commonly occurring names and addresses can be replaced with standard spellings using dictionaries. Other errors that need to be addressed are English transliterations of foreign names; the use of initials, truncations and abbreviations; and swapping of surnames and forenames.
- Strings might contain extra words such as 'Mr', 'Mrs', 'Dr', 'Jr' or directional 'East' or 'West'. These redundant words are typically removed by a direct linkage to a dictionary containing a list of such redundant words.
- All missing and miscoded data need to have the same definition and notation across the datasets. Missing values have to be consistently labelled as such.
- All files have to have standardised formats for each of the variables to be used for matching, for example, coding of dates should be the consistent across datasets.

- All matching variables must have the same characteristics, field length and coding status across datasets.

3.1.2 Phonetic Codes and String Comparators

To compensate for errors in strings, probabilistic data linkage can make use of phonetic codes and string comparators. Phonetic codes cope with spelling errors, for example 'Reid' and 'Reed', would contain the same phonetic code, as they sound the same and hence would be considered an agreement if they were compared on their phonetic code. The most commonly used phonetic code is Soundex because it exists in many statistical packages. However, for foreign names the code is less satisfactory, since it ignores vowel sounds. There are variations of Soundex that have been developed in different countries. The use of Soundex may, however, cause pairs to agree in a string when in fact they are very different. More robust phonetic codes have been developed. One such code is the New York State Identification and Intelligence System (NYSIIS) code. This code retains information about the position of vowels by converting most vowels to the letter 'A', and it replaces consonants with other, phonetically similar, letters.

String comparator metrics are another way to deal with typographical errors by accounting for deletions, insertions and transpositions (where a letter is moved one position to the left or right) in strings. A string comparator $\Phi_{(S_1,S_2)}$ is a metric between 0 and 1 where 1 denotes a perfect agreement and 0 denotes a perfect disagreement. Jaro (1989) introduced a string comparator that has been shown to be robust when used for data linkage of individuals and is commonly used when linking first and last names. The algorithm is based on the lengths of the two strings denoted by *str_length1* and *str_length2*, the number of common characters across the two strings (the common letter must be within half of the length of the smaller string), denoted *#common*, and the number of transpositions, denoted *#transpositions*. It is defined as follows:

$$\Phi_{(S_1,S_2)} = \frac{1}{3}\left[\frac{\#common}{str\_length1} + \frac{\#common}{str\_length2} + \left(1 - \frac{1}{2}\left(\frac{\#transpositions}{\#common}\right)\right)\right] \quad (1)$$

Winkler (1990) found that fewer errors are made at the beginning of the string than at the end of the string and hence has enhanced the Jaro string comparator by introducing weights. This is known as the Jaro-Winkler string comparator.

Another commonly used string comparator is bigrams. These are typically used in privacy-preserving data linkage where strings are anonymised via computer science functions (see Sect. 4.2). A bigram is two consecutive characters in a string. For example, bigrams in the word 'bigram' are 'bi', 'ig', 'gr', 'ra' and 'am'. The string comparator is defined as the proportion of two character sub-strings in common between the two strings, where the denominator is the average number of sub-strings.

3.1.3 Blocking Variables

In addition to selecting matching variables, there is a need to determine blocking variables. A blocking variable aims to reduce the search space between two datasets by avoiding the comparison of record pairs that are least likely to be matches. For example, if both datasets have 10,000 individuals and a one-to-one match is to be carried out, this results in 100 million pairs to compare. The search space can be dramatically reduced by forming pairs for comparison only among those with the potential to be matches, such as those with a common geographical area. For example, the first three digits of a postcode can be used as a blocking variable. Record pairs are brought together only if they agree (exactly) on the blocking variable. This use of a deterministic matching approach to assist in the probabilistic data linkage greatly reduces the computational burden. However, blocking variables must be as error-free as possible or potential matches can be missed.

The typical approach, especially for one-to-one matching, is to carry out the probabilistic data linkage sequentially, starting with a restrictive deterministic matching on the blocking variable and forming all record pairs for comparison. The pairs are compared, and matches are determined following the method described below. Once matches are determined, they are set aside and a second linkage is carried out through the residual datasets, where the blocking variable can now be less restricted. This is carried out multiple times until the residual datasets are small enough that no blocking variable is needed.

The blocking variables must be small enough to avoid too many unproductive comparisons but large enough to prevent records for the same entity spilling over into adjacent blocks and so failing to compare possible true matches. Therefore, different blocking variables should be used for the iterative passes through the datasets. For example, one might block on postcode and surname, carry out the data linkage and place matches aside and then match residual datasets using a different blocking criteria, such as year of birth or initial of first name and so on.

Once the blocking variable and matching variables are determined for the current iteration, both datasets need to be blocked and sorted and all possible pairs generated. This can easily be done through database Structured Query Language (SQL) commands used for relational database management systems.

3.2 Linkage Stage

3.2.1 Parameters of Data Linkage

In probabilistic data linkage, a frequency analysis of data values is carried out to calculate a weight or score for each matching variable, which indicates how likely it is that they refer to the same entity. Uncommon value agreements should give stronger evidence of linkage. Large weights assigned to matching variables are expected when there is a correct match and small weights assigned to matching

variables when there is no match. Note that there is still a positive, albeit small, weight even for an incorrect match, due to the potential for errors in the matching variable. The weight is a ratio of two frequencies:

- Number of agreements of the value of the matching variable in record pairs that represent that same entity (true match)
- Number of agreements of the value of the matching variable in record pairs that do not represent the same entity

This original framework for data linkage was developed by Newcombe et al. (1959) and formalised into the well-known probabilistic framework described in Fellegi and Sunter (1969), referred to hereafter as F&S.

There are three key parameters for probabilistic data linkage, which are represented by the following concepts:

- The quality of the data
- The chance that the values of a matching variable will randomly agree
- The ultimate number of true matches that exist in the database

The quality of the data is represented by the numerator of the above ratio and is denoted as the *m*-probability in the F&S framework: the probability that a matching variable agrees given that the pair is a true match. This is the degree to which the information contained for a matching variable is accurate and stable across time. Data entry errors, missing data or false dates diminish accuracy and produce low-quality data.

The discriminating power of the matching variable is represented by the denominator of the above ratio and is denoted as the *u*-probability in the F&S framework: the probability that a matching variable agrees given that the pair is not a true match. This is similar to the situation where a matching variable will randomly agree across a pair regardless of whether it is a true match or not a true match and is approximately equal to the inverse of the number of values of the matching variable. For example, gender would be expected to randomly agree 50% of the time between pairs and hence does not have high discriminating power.

The third parameter is the ultimate number of true matches or the marginal probability of a correct match. Although this parameter is not explicit in the above ratio, it is essential that there is a sufficiently high proportion of true matches to ensure a successful data linkage.

3.2.2 Basic Concepts in Probabilistic Data Linkage

As mentioned, probabilistic data linkage relies on calculating weights or scores for each matching variable, based on a frequency analysis of the number of agreements as well as disagreements in pairs of records. In the simplest approach, probabilistic data linkage requires some preliminary matching to have been carried out on a similar application. For example, to estimate census undercounts from a post-enumeration survey, the linkage between the survey and the census relies largely

on parameters derived from the previous census. Alternatively, a gold standard matched test dataset can be constructed from a small portion of the datasets that have been carefully verified and checked. From this test dataset, the probabilities are calculated that two records will agree on a matching variable in truly matched pairs compared with the probability that records will agree on non-matched pairs or simply by chance. In other words, how likely is it that the variables that agree between a record pair would have done so by chance if the pair was not correctly matched (the *u*-probability)? This is compared with how likely the agreement would be in correctly matched record pairs (the *m*-probability). This criterion therefore determines good matching variables, i.e. the agreement between variables should be more typical of correctly matched pairs than of those that might have occurred by chance in unrelated records. Section 3.2.4 describes the expectation-maximisation (EM) algorithm for estimating *m*- and *u*-probabilities without the need for test data.

In formal notation in the F&S framework:

For two datasets A and B, let the records in each dataset be denoted $a \in A$, $b \in B$ and the set of all possible matches $A \times B = \{(a,b); a \in A, b \in B\}$. Let $\alpha(a)$ represent the matching variables for entity a in file A and similarly $\beta(b)$ for entity b in file B. The aim is to determine a set of matches $M = \{(\alpha(a), \beta(b)) | \, a = b\}$ and a set of non-matches $NM = \{(\alpha(a), \beta(b)) | \, a \neq b\}$. To develop the decision rule, it is necessary to define a comparison space $C : \alpha(a) \times \beta(b) \rightarrow \Gamma$. This comparison space is composed of a comparison vector $\gamma \in \Gamma$ that represents an agreement pattern (typically 1 for agree and 0 for disagree) for each matching variable. As an example of an agreement pattern for pair j with three matching variables $\gamma^j = \left(\gamma_1^j, \gamma_2^j, \gamma_3^j\right)$, let $\gamma_1^j = 1$ if pair j agrees on last name and 0 otherwise, $\gamma_2^j = 1$ if pair j agrees on first name and 0 otherwise, and $\gamma_3^j = 1$ if pair j agrees on street name and 0 otherwise. One such agreement pattern might be $\gamma^j = (1,0,1)$: agree on last name, disagree on first name and agree on street name. In fact, for three matching variables and for a simple agree/disagree {1,0} pattern, the comparison space would contain eight possible agreement patterns. Agreement patterns can also be more complex, using string comparators, for example, $\gamma^j = (0.66, 0, 0.80)$.

The *m*-probability is now formally defined as the conditional probability that a record pair j has an agreement pattern γ^j given that it is a match (M), denoted as $m = P(\gamma^j | \, M)$, and the *u*-probability as the conditional probability that a record pair j has an agreement pattern γ^j given that it is not a match (NM), denoted as $u = P(\gamma^j | \, NM)$. Finally, let $P(M)$ be the marginal probability of a correct match.

The probability of interest is the match probability given an agreement pattern γ: $P(M | \, \gamma^j)$. According to Bayes' theorem, this is the posterior probability calculated as follows:

$$P\left(M|\gamma^j\right) = \frac{P\left(\gamma^j|M\right)P(M)}{P\left(\gamma^j\right)} = \frac{P\left(\gamma^j|M\right)P(M)}{P\left(\gamma^j|M\right)P(M)+P\left(\gamma^j|NM\right)(1-P(M))} = \frac{1}{1+\frac{P\left(\gamma^j|NM\right)(1-P(M))}{P\left(\gamma^j|M\right)P(M)}} \tag{2}$$

The agreement (likelihood) ratio $R\left(\gamma^j\right) = \frac{P\left(\gamma^j|M\right)}{P\left(\gamma^j|NM\right)}$ is defined as the test statistic (overall score) for record pair j, since maximising the likelihood ratio is the same as maximising the posterior probability of $P(M|\,\gamma^j)$. Therefore, one can simply order the likelihood ratios $R(\gamma^j)$ and choose an upper cutoff W^+ and a lower cutoff W^- for determining the correct matches and correct non-matches. The linkage rule $F : \Gamma \rightarrow \{M, C, NM\}$ maps a record pair j comparison value to a set of three classes—matches (M), non-matches (NM) and a set of undecided cases for manual clerical review (C)—defined as follows:

$$F : \begin{cases} \gamma^j \in M \quad if \quad R\left(\gamma^j\right) \geq W^+ \\ \gamma^j \in NM \quad if \quad R\left(\gamma^j\right) \leq W^- \\ \gamma^j \in C \quad otherwise \end{cases} \tag{3}$$

The F&S framework assumes conditional independence across matching variables. This means that the errors associated with one matching variable are independent of the errors associated with another matching variable. Under conditional independence the m- and u-probabilities can be decomposed as follows: $P\left(\gamma^j|M\right) = P\left(\gamma_1^j|M\right) \times P\left(\gamma_2^j|M\right) \times \cdots \times P\left(\gamma_k^j|M\right)$ and $P\left(\gamma^j|NM\right) = P\left(\gamma_1^j NM\right) \times P\left(\gamma_2^j|NM\right) \times \cdots \times P\left(\gamma_k^j|NM\right)$. The likelihood ratio for record pair j becomes:

$$R\left(\gamma^j\right) = \frac{P\left(\gamma^j|M\right)}{P\left(\gamma^j|NM\right)} = \frac{P\left(\gamma_1^j|M\right) \times P\left(\gamma_2^j|M\right) \times \cdots \times P\left(\gamma_k^j|M\right)}{P\left(\gamma_1^j|NM\right) \times P\left(\gamma_2^j|NM\right) \times \cdots \times P\left(\gamma_k^j|NM\right)}$$

Taking the log transformation, the overall score based on the likelihood ratio for record pair j is the sum:

$$\begin{aligned} \log\left[R\left(\gamma^j\right)\right] = &\log\left(\frac{P\left(\gamma_1^j|M\right)}{P\left(\gamma_1^j|NM\right)}\right) + \log\left(\frac{P\left(\gamma_2^j|M\right)}{P\left(\gamma_2^j|NM\right)}\right) \\ &+ \cdots + \log\left(\frac{P\left(\gamma_k^j|M\right)}{P\left(\gamma_k^j|NM\right)}\right) \end{aligned} \tag{4}$$

Note that any log can be taken for the transformation and that here the natural log is used.

For example, assume from a previous linkage that the following m- and u-probabilities were obtained:

P(agree on characteristic x|M) = 0.9 if x = first name, last name, year of birth and 0.8 if x = house number, street name, gender

P(agree on characteristic x|NM) = 0.05 if x = first name, last name, year of birth and 0.1 if x = house number, street name, gender

Assume the following record pair *j* is to be examined:

Name	Address	Age	Gender
Barbara Jones	439 Elm St	1968	M
Barbara Jones	435 Elm St	1969	F

The agreement vector is γ^j = (agree first name, agree last name, disagree house number, agree street name, disagree year of birth, disagree gender) = (1, 1, 0, 1, 0). When there is a disagreement in a matching variable *k*, the complement of the likelihood ratio or the disagreement ratio is calculated as $\frac{\left[1-P\left(\gamma_k^j|M\right)\right]}{\left[1-P\left(\gamma_k^j|NM\right)\right]}$. The overall score for record pair *j* with the agreement vector (1, 1, 0, 1, 0), and based on the likelihood ratio of each matching variable, is:

$$\begin{aligned}\log\left(R\left(\gamma^j\right)\right) = &\log(0.9/0.05) + \log(0.9/0.05) + \log((1-0.8)/(1-0.1)) \\ &+ \log(0.8/0.1) + \log((1-0.9)/(1-0.05)) \\ &+ \log((1-0.8)/(1-0.1)) = 1.129\end{aligned}$$

Similarly, the overall scores are calculated for all record pairs.

While the *m*-probability represents the quality of the matching variable and is not dependent on the actual value of the matching variable, this is not the case for the *u*-probability. The *u*-probability represents the discriminating power of the matching variable, and hence rare values should provide more weight to the overall score than common values. In addition, since the number of non-matches is very large compared with the number of matches when comparing all possible pairs within blocks, the *u*-probability $P(\gamma|NM)$ is often approximated by the marginal probability $P(\gamma)$. For example, the *u*-probability of month of birth is often taken as 1/12 and gender as 1/2. Therefore, in many small-scale applications, the *u*-probability is calculated as the proportion of value states of the matching variable in a large dataset or across all possible pairs. However, the calculation of the *m*-probability needs good test data or an approach such as the EM algorithm described in Sect. 3.2.4, since this is calculated as the rate of error among known matches.

The likelihood ratio can be modified to take into account a string comparator. A common approach when using the simple agree/disagree {1,0} comparison vector of the F&S framework is by interpolation. Assume matching variable *k* is first or last name. The likelihood ratio is modified as follows:

$$R\left(\gamma_k^j\right) = \Phi^j_{(S_1,S_2)}\frac{P\left(\gamma_k^j|M\right)}{P\left(\gamma_k^j|NM\right)} + \left(1-\Phi^j_{(S_1,S_2)}\right)\frac{1-P\left(\gamma_k^j|M\right)}{1-P\left(\gamma_k^j|NM\right)} \quad (5)$$

One can see that if there is a perfect agreement and $\Phi^{j}_{(S_1,S_2)} = 1$, then we obtain the original agreement likelihood ratio, and when $\Phi^{j}_{(S_1,S_2)} = 0$ we obtain the disagreement likelihood ratio. Intermediary values are obtained for the likelihood ratio under partial agreements. Finally, on the new likelihood ratio, the log transformation is taken and added to the likelihoods ratios of the other matching variables.

For missing values, one might consider taking a comparator $\Phi_{(S_1,S_2)}$ of $1/k$ where k is the number of categories. For example, a missing value in gender could have a string comparator of $\Phi_{(S_1,S_2)} = 1/2$.

For a quantitative variable such as year of birth, one might want to provide more of an agreement if the difference in year of birth is 1 or 2 years compared with a difference in year of birth of more than 3 years. A possible string comparator is:

$$\Phi_{(S_1,S_2)} = \begin{cases} \exp(-|birth\ \ year1 - birth\ \ year2|/3)\ if & |\ birth\ \ year1 - birth\ \ year2\ | < 3 \\ 0 & otherwise \end{cases}$$

which obtains a value of 1 if *birth year 1* = *birth year 2*, a value of 0.717 if there is a difference in 1 year, 0.513 if there is a different in 2 years and 0 otherwise.

3.2.3 Setting Thresholds

Based on the observed values of the comparison vector between a record pair, F&S consider the (log) ratio of probabilities in Eq. (4). A decision rule is given by thresholds W^{+} and W^{-} as shown in Eq. (3) which are determined by a priori error bounds. The errors are defined as Type 1 (false matches) and Type 2 (false non-matches) errors. These error bounds are preset by the data linker and should be very small. Generally, the Type 2 error bound is larger than the Type 1 error bound because data linkage is typically an iterative process, with multiple passes through the datasets, and hence missed matched pairs may be found in subsequent passes through the record pairs using different blocking variables. On the other hand, in a Type 1 error, paired records may be obtained that are erroneously declared matches, which can cause severe bias in statistical analysis on the matched dataset. The undetermined record pairs (those record pairs between the upper and lower cutoff thresholds) are sent for clerical review to determine their match status. Fellegi and Sunter (1969) show that, given the error bounds, the decision rule in Eq. (3) is optimal and minimises the number of pairs that need to be clerically reviewed.

The thresholds can be determined by a test dataset where the true match status is known. The empirical distribution of the overall scores from the matches and non-matches are used separately to determine the cutoff thresholds. Before calculating the empirical distribution, the overall scores (likelihood ratios) for each record pair j should be transformed into a probability according to the following transformation:

$p_j = \frac{\exp(W_0+W_j)}{\exp(W_0+W_j)+1}$ where $W_0 = \log\left(\frac{E}{A\times B-E}\right)$ and $A \times B$ is the total number of pairs, E is the expected number of matches and $W_j = \log\left[R(\gamma^j)\right]$ the overall score calculated in Eq. (4).

Based on the transformed overall scores p_j, we calculate the cumulative empirical distribution of the matches and find the threshold W^- that corresponds to the predetermined Type 2 error bound in the lower tail. We then calculate the empirical distribution of the non-matches and find the threshold W^+ that corresponds to the predetermined Type 1 error bound in the upper tail. All pairs above W^+ are declared matches, all pairs below W^- are declared non-matches and those in between are sent for clerical review, as shown in the decision rule in Eq. (3). Often, the extent of clerical review is determined by available resources. The costs can also be incorporated into the decision rule in Eq. (3) (Pepe 2003).

3.2.4 Expectation-Maximisation Algorithm for Estimating Parameters

Fellegi and Sunter (1969) considered the decomposition of the probability of agreement for record pair j under the simple agree/disagree {1,0} comparison patterns:

$$P\left(\gamma^j\right) = P\left(\gamma^j|M\right)P(M) + P\left(\gamma^j|NM\right)(1-P(M)) \tag{6}$$

The left-hand side of Eq. (6) obtains the proportion of the agreement patterns across all possible pairs. For example, for three matching variables, there would be 8 ($= 2^3$) possible agreement patterns and hence 8 equations; although since probabilities must sum to 1, the 8th equation is redundant. These probabilities can be used to solve for the probabilities on the right-hand side of Eq. (6). Assuming a simple agree/disagree {1,0} pattern for each matching variable, the m-probability for a matching variable k in record pair j is distributed according to the Bernoulli distribution $P\left(\gamma_k^j|M\right) = m_k^{\gamma_k^j}(1-m_k)^{1-\gamma_k^j}$ and under the assumption of conditional independence across all matching variables: $P\left(\gamma^j|M\right) = \prod_k m_k^{\gamma_k^j}(1-m_k)^{1-\gamma_k^j}$. Similarly, for the u-probability: $P\left(\gamma^j|NM\right) = \prod_k u_k^{\gamma_k^j}(1-u_k)^{1-\gamma_k^j}$. Therefore, for each matching variable k, there are two unknown probabilities, m_k and u_k, as well as the overall match probability, $P(M)$. With three matching variables, seven unknown parameters are obtained. Fellegi and Sunter (1969) showed that, by using information from the frequencies of the agreement patterns on the left-hand side of Eq. (6), one can estimate these unknown probabilities on the right-hand side of Eq. (6).

Data linkage will typically have more than three matching variables; thus the aim of the EM algorithm is to find the best solution. In the E-step, the indicator value is

estimated for the true match status denoted by $g_m^j = 1$ if record pair j represents the same entity (set M) or 0 otherwise and $g_u^j = 1$ if record pair j does not represent the same entity (set NM) or 0 otherwise. Applying Bayes' theorem for the simple case of agree/disagree {1,0} agreement pattern and starting with initialising values for the probability of a match (denoted $\widehat{p}$) and m- and u-probabilities for each matching variable, the estimates of the indicator values for the jth record pair are:

$$\widehat{g}_m^j = \frac{\widehat{p} \prod_k \widehat{m}_k^{\gamma_k^j} (1 - \widehat{m}_k)^{1-\gamma_k^j}}{\widehat{p} \prod_k \widehat{m}_k^{\gamma_k^j} (1 - \widehat{m}_k)^{1-\gamma_k^j} + (1 - \widehat{p}) \prod_k \widehat{u}_k^{\gamma_k^j} (1 - \widehat{u}_k)^{1-\gamma_k^j}}$$

and

$$\widehat{g}_u^j = \frac{(1 - \widehat{p}) \prod_k \widehat{u}_k^{\gamma_k^j} (1 - \widehat{u}_k)^{1-\gamma_k^j}}{\widehat{p} \prod_k \widehat{m}_k^{\gamma_k^j} (1 - \widehat{m}_k)^{1-\gamma_k^j} + (1 - \widehat{p}) \prod_k \widehat{u}_k^{\gamma_k^j} (1 - \widehat{u}_k)^{1-\gamma_k^j}}$$

In the M-step, the values of the three probabilities are updated, m-probability, u-probability and the proportion of matched pairs $\widehat{p}$, as follows: $\widehat{m}_k = \frac{\sum_j g_m^j \gamma_k^j}{\sum_j g_m^j}$, $\widehat{u}_k = \frac{\sum_j g_u^j \gamma_k^j}{\sum_j g_u^j}$ and $\widehat{p} = \frac{\sum_j g_m^j}{R}$, where R is the number of record pairs.

These new estimates can be replaced in the E-step and iterated until convergence, i.e. the difference between the probabilities at iteration $t - 1$ and iteration t is below a small threshold. One can also plug in the u-probabilities if they are value-specific and known from a large database and use the EM algorithm to estimate the m-probabilities and the overall match probability $\widehat{p}$.

3.3 Post-linkage Stage and Evaluation Measures

After the data linkage process, it is necessary to carry out checks for errors in the match status of the matched and non-matched dataset. A small random sample is drawn from the set of matches and the set of non-matches and the accuracy of the match status is verified, particularly for those record pairs near the threshold cutoff values. These checks allow accurate estimation of the Type 1 and Type 2 errors.

In terms of classic decision theory, the decision matrix for data linkage is presented in Table 1.

The probability of a Type 1 error is the proportion of falsely linked non-matches out of the total number of non-matches. The specificity is the compliment and represents the correctly non-linked matches out of the total number of non-matches.

Table 1 Decision matrix for data linkage

Decision	No match (null hypothesis)	Match (alternative hypothesis)
No link (do not reject null)	Ok (true negative)	Type 2 error (false negative)
Link (reject null)	Type 1 error (false positive)	Ok (true positive)

The probability of a Type 2 error is the proportion of incorrectly non-linked pairs out of the total number of true matches. The compliment is the power of the test, which is the proportion of correctly paired matches out of the total number of true matches. This is also known as the sensitivity. In the data linkage literature, it is also known as recall. Another measure found in the data linkage literature is precision, which is defined as the number of correctly linked matches out of the total number of linked pairs.

In summary:

- Sensitivity/recall—correctly matched pairs out of all true matches
- Specificity—correctly not linking non-matches out of all true non-matches
- Precision—correctly matched pairs out of all possible decisive links

It is important to disseminate these measures to users of the linked data to enable them to understand the quality of the linkage and be able to compensate for linkage errors in statistical analysis and inference (see Sect. 5).

3.4 Constraints on Matching

The essence of a probabilistic data linkage is iterating passes of the datasets in which blocking variables (must match exactly) and matching variables (used to compute the agreement scores) change roles. Blocking variables reduce the computational burden but increase the false non-match rate. Matching variables increase the computational burden and manage the trade-off between false match and false non-match errors. Multiple passes through the pairs are carried out, interchanging blocking and matching variables. As records are linked, they are removed from the input files, and therefore one can use fewer blocking variables to avoid the chance of false non-matches.

In many cases, it may be necessary to match hierarchical datasets, for example, all individuals within households. The characteristics associated with households, such as street name and street number, may overwhelm the characteristics associated with individuals and diminish the power of the statistical test. This is solved by first matching households based on household variables and then matching individuals within households. This partitions the set of pairs as matches within households, non-matches within households and non-matches outside of households. All non-matches are collapsed into one class for the next iteration through the record pairs.

Another requirement is to ensure that matching variables are not highly correlated with each other, since this diminishes the discriminating power of the variable.

For example, age and year of birth should not be used together as matching variables for data linkage.

4 Recent Advances

Two areas that have shown much development in recent years are clustering algorithms for linkage across multiple databases, specifically for deduplication, and privacy-preserving record linkage. More details are provided below.

4.1 Indexing and Blocking

Traditionally, blocking variables partition records based on an exact match to values of a variable such as postcode or first initial of first name (see Sect. 3.1.3), and the data linkage is carried out iteratively, exchanging blocking and matching variables. More formally, this is known as indexing, which is the procedure to reduce the search space of record pairs that are unlikely to contain true matches. Multipurpose indexing can be carried out by performing multiple blocking passes and using different blocking variables and then taking the union of all the retained pairs. Other approaches include a neighbourhood approach, where records are sorted on each database and a sliding window of fixed size is used to define the blocks, and canopy clustering, where a distance metric (e.g. similarity score) is calculated between the blocking variables and records are inserted into one or more clusters. Then, each cluster becomes a block from which record pairs are produced. Filtering discards any pairs not initially excluded by the blocking variables but which are still unlikely to contain a true match. These techniques and others are described in Christen (2012) and Murray (2016). Steorts et al. (2016) and Sadinle (2017) have also applied innovative methods for indexing and deduplication in multiple databases using clustering techniques. This has led to new representations of the data linkage problem within a Bayesian framework.

4.2 Privacy-preserving Record Linkage

In privacy-preserving record linkage, the matching variables are encrypted using computer science techniques instead of coarsened, perturbed or deleted, as is the usual practice in statistical disclosure control. The encryption of matching variables takes the form of ‘hashing’, where the strings are first split into bigrams and then hash functions are used to encrypt each bigram. Another approach is the use of Bloom filters, where strings are encoded into a data structure defined as a bit string (an array of 0s and 1s) that can be represented as an integer in its binary form and used to test whether or not an element is a member of the set (Schnell et

al. 2009). Even if matching variables are encrypted, they can be used in exact matching. Furthermore, similarity scores can be used as string comparators; the most commonly used is the Jaccard score. For the hashed bigrams, the Jaccard score is the ratio of the exact matching bigrams divided by the total number of bigrams. For the bloom filter, the Jaccard score is the ratio of the common bits in the string divided by the total number of bits.

As a relatively new area of research, privacy-preserving record linkage is making the crossover from computer science into the statistics community. There is still much work to be done to prove that the method is viable and 'fit for purpose'. Smith and Shlomo (2014) propose data linkage within and across data archives with data custodians allowing the encryption of matching variables based on a common seed. Users can then request a probabilistic data linkage based on the F&S approach through an interface.

One of the drawbacks of privacy-preserving record linkage is that clerical review cannot be carried out except by a trusted third party, who would have access to the original strings and the keys for the encryption.

5 Analysis of Linked Data

Research into the analysis of linked data that are subject to linkage errors has recently gained traction. Earlier work by Scheuren and Winkler (1993, 1997) followed by Lahiri and Larsen (2005) and Chambers (2009) has dealt with the problem of linkage errors in regression modelling. Assume a one-to-one match on two datasets of the same size where a variable X on file A is linked to a variable Y from file B. C is defined as the permutation matrix of {0,1} representing the data linkage. For example:

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_3 & y_1 \\ x_1 & y_2 \\ x_2 & y_3 \end{pmatrix}$$

Consider the regression equation $Y = CX\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 I_n)$. A naive and biased estimate of β is $\widehat{\beta}_N = \left(X'C'CX\right)^{-1} X'C'Y$. If both X and Y are written in their design matrix form, the naive contingency table can be written as $X'C'Y$.

We assume that C is a random permutation matrix whose distribution is dependent on the parametric estimates of the record linkage denoted by ψ and on X. Define $Q = E(C|X, \psi)$ as a probability error matrix as follows:

$$\begin{cases} Y_i^* = Y_i & with \quad probability \quad q_{ii} \\ Y_i^* = Y_j & with \quad probability \quad q_{ij}\, i \neq j \end{cases}$$

where Y_i^* is the linked record and Y_i is the true record.

Lahiri and Larsen (2005) propose an unbiased estimator for the regression parameter β: $\widehat{\beta}_{LL} = \left(X'Q'QX\right)^{-1}X'Q'Y^*$. For the contingency table, an unbiased estimator is $X'\,Q^{-1}\,Y^*$. An additional estimator proposed for the contingency table is $X'\,Q'\,Y^*$, and although this is a biased estimator, it will generally have smaller mean square error than the unbiased estimator. This is the subject for further research.

How to obtain the error matrix Q is still a subject of open debate. Lahiri and Larsen (2005) suggest using the matching probabilities, themselves derived from the linkage process to define the Q matrix. Chambers (2009) mentions that researchers analysing the data would probably not have access to all the matching probabilities and proposes working with the data linkers to estimate linkage errors using small random samples in a post-linkage step.

Chambers (2009) defines the exchangeable linkage error model for estimating the Q matrix under the following assumptions: one-to-one match and no missed links; data linkage errors occur only within distinct blocks m, $m = 1,2 \ldots M$, and each block is of size n_m; within each block m, the linkage is non-informative, i.e. the probability of a correct match is the same for all records in the block; and it is equally likely that any two records could in fact be the correct match.

Based on these criteria, the Q matrix is defined as follows:

Let $\Pr(correct\ \ link) = P(C_{m(i,i)} = 1) = q_m$ and $\Pr(not\ \ correct\ \ link) = P(C_{m(i,j)} = 1) = (1 - q_m)/(n_m - 1)$,

where Q is a block diagonal matrix of m blocks of size n_m and in each block q_m is on the diagonal and $(1 - q_m)/(n_m - 1)$ is on the off-diagonal. Note that the row sums to 1, as is the requirement for a probability error matrix. Chambers (2009) also proposed a modification that adapts for false non-matches, which would not appear in the dataset.

In another approach, Goldstein et al. (2012) treat the equivocal links (those links that are not an exact match) as missing values. They then impute the values under two approaches: standard multiple imputation and extended multiple imputation, where the posterior distribution is adjusted by priors defined by the matching probabilities. Under the non-informative linkage assumption (similar to the exchangeable linkage error model) and assuming no model misspecification for the multiple imputation of the equivocal links, the authors achieve good results for reducing bias caused by linkage errors.

Further research is ongoing into other types of regression modelling, the relaxation of the restrictions of the exchangeable linkage error model and also the use of calibration to compensate for missed links. One significant problem that remains is specifying the error probabilities in the Q matrix. It may be possible under certain conditions to simulate the matching process based on the matching parameters and estimate error rates through a bootstrap procedure, as described in Winglee et al. (2005) and Chipperfield and Chambers (2015).

References

Chambers R (2009) Regression analysis of probability-linked data, Official Statistics Research Series 4. Statistics New Zealand, Wellington http://www3.stats.govt.nz/statisphere/Official_Statistics_Research_Series/Regression_Analysis_of_Probability-Linked_Data.pdf. Accessed 2018

Chipperfield JO, Chambers R (2015) Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. J Off Stat 31(3):397–414

Christen P (2012) Data matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Springer-Verlag, Berlin

D'Orazio M, Di Zio M, Scanu M (2006) Statistical matching. Wiley, Chichester

Fellegi IP, Sunter AB (1969) A theory for record linkage. J Am Stat Assoc 64:1183–1210

Goldstein H, Harron K, Wade A (2012) The analysis of record-linked data using multiple imputation with data value priors. Stat Med 31(28):3481–3493

Jaro MA (1989) Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. J Am Stat Assoc 84(406):414–420

Lahiri P, Larsen M (2005) Regression analysis with linked data. J Am Stat Assoc 100:222–230

Murray JS (2016) Probabilistic record linkage and deduplication after indexing, blocking, and filtering. J Priv Confid 7(1):3–24

Newcombe HB, Kennedy JM, Axford SJ et al (1959) Automatic linkage of vital records. Science 130(3381):954–959

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, New York

Sadinle M (2017) Bayesian estimation of bipartite matchings for record linkage. J Am Stat Assoc 112(518):600–612

Scheuren F, Winkler WE (1993) Regression analysis of data files that are computer matched – part I. Surv Methodol 19:39–58

Scheuren F, Winkler WE (1997) Regression analysis of data files that are computer matched – part II. Surv Methodol 23:157–165

Schnell R, Bachteler T, Reiher J (2009) Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak 9:41

Smith D, Shlomo N (2014) Record linkage approaches for dynamic database integration: privacy preserving probabilistic record linkage, deliverable 11.1 for WP11, Data Without Boundaries. http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2014-01-Data_without_Boundaries_Report.pdf. Accessed 2018

Steorts RC, Hall R, Fienberg S (2016) A Bayesian approach to graphical record linkage and deduplication. J Am Stat Assoc 111(516):1660–1672

Winglee M, Valliant R, Scheuren F (2005) A case study in record linkage. Surv Methodol 31(1):3–12

Winkler WE (1990) String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In: JSM proceedings, Survey research methods section. American Statistical Association, Alexandria, VA, pp 354–359 Retrieved from http://ww2.amstat.org/sections/srms/Proceedings/. Accessed 2018

Natalie Shlomo (BSc, Mathematics and Statistics, Hebrew University; MA, Statistics, Hebrew University; PhD, Statistics, Hebrew University) is Professor of Social Statistics in the School of Social Sciences at the University of Manchester. Her area of interest is in survey statistics covering

survey design and estimation, record linkage, statistical disclosure control, statistical data editing and imputation, non-response analysis and adjustments, adaptive survey designs, quality indicators for survey representativeness and small area estimation.