A unified approach to inference with linked data

Rachel Anderson*

This Version: September 9, 2019

Abstract

This paper studies the joint effects of matching algorithms and estimation procedures for linked data on the quality of the estimates that they produce. Specifically, I compare how different combinations of record linkage and estimation methods perform across a variety of data generating processes and research tasks (i.e. linking the dependent variable, independent variable, sample selection, GMM, linear regression) in order to develop a unified approach for doing empirical work with linked data.

1 Introduction

In empirical microeconomics, identifying a common set of individuals appearing in two or more datasets is often complicated by the absence of unique identifying variables. For example, economic historians frequently use historical U.S. data that lack Social Security Numbers, so that matching observations across datasets requires using characteristics such as name and reported age[1]. Yet the presence of common names, along with transcription and enumeration errors, age misreporting, mortality, under-enumeration, and migration between

*Mailing Address: Department of Economics, Julis Romo Rabinowitz Building, Princeton, NJ 08544. Email: rachelsa@Princeton.edu. This project received funding from the NIH (Grant 1 R01 HD077227-01).

[1]Depending on the data source, additional variables such as middle name or initial, birthplace, parents' name and birthplace, are sometimes available.

census years, makes performing an error-free one-to-one merge – the most common data pre-processing step in such analyses – virtually impossible (Abramitzky et al., 2019a).

For example, Aizer et al. (2016) link children listed on Mothers' Pension program welfare applications from 1911-1935 with records from the Social Security Death Master File, recorded between 1965 and 2012. Using an automated matching procedure that links individuals based on their name and date of birth, the authors match 48 percent of individuals to a unique death record, and 4 percent to multiple possible records, leaving 48 percent of individuals unmatched.[2]

This example illustrates how inference using matched data requires deciding how to variably treat observations with unique, multiple, or zero matches. Aizer et al. (2016) use methods from Anderson et al. (2019), which assume that each of the linked death record is equally likely to be the true match; however other authors have approached the same problem by generating a "composite match" equal to the average of the linked observations (Bleakley and Ferrie, 2016), and constructing bounds on the parameter of interest using different configurations of matched data (Nix and Qian, 2015). Alternatively, the authors could have used probabilistic record linkage methods to link the data, which would have allowed them to use the OLS bias correction from Lahiri and Larsen (2005) or the prior-informed imputation strategy for missing records proposed by Goldstein et al. (2012).

While there are a number of papers that compare the performance of different matching algorithms (Bailey et al., 2017; Abramitzky et al., 2018), as well as the performance of different estimation methods for linked data (Harron et al., 2014), there are few (if any) that study the joint impact of matching and estimation on inference. The only exception is Scheuren and Winkler (1997), who suggest iterating between estimation, imputation, and record linkage steps when the goal is linear regression.

[2]The authors estimate that at least 32 percent of individuals in the Mothers' Pension program data died before 1965, and therefore should have no match in the 1965-2012 data.

however, the authors describe how to construct more efficient estimators if additional information about match quality is available. Specifically, if the researcher can estimate the probability that each individual-outcome pair is a true match, then this knowledge can be used to achieve a reduction in mean-squared error. Such probabilities are outputted by probabilistic record linkage procedures, first developed by Fellegi and Sunter (1969) in the statistics literature, but only recently applied to economics Abramitzky et al. (2018). Hence, any discussion of best practices for using linked data should address how to choose matching algorithms and estimation procedures jointly.

The goal of this paper is therefore to study the effects of different combinations of matching algorithms and estimation procedures for linked data on the quality of the estimates that they produce. First, I will compare how different matching algorithms perform in terms of the representativeness of the matched data they produce and their tolerance for type I and type II errors. Next, with multiple matched versions of the data in hand, I will compute point estimates and confidence intervals for the same parameter of interest using methods that vary by whether they allow for multiple matches, incorporate the matching probabilities, and are likelihood-based in their approach. In total, I will perform the above analysis twice – with simulated data and with real data that the simulated data are generated to imitate.

To the best of my knowledge, how data pre-processing impacts subsequent inference in economics research is not well understood. This paper adds to a recent series of papers by Bailey et al. (2017) and Abramitzky et al. (2018, 2019a), which evaluate the performance of common matching algorithms for historical data in a variety of real and simulated data settings, and offer informal discussions about the impact of matching on subsequent inference. This paper pushes these ideas further, by measuring the *joint* effects of matching and estimation; with a focus on developing estimation techniques that incorporate information from the matching process to improve efficiency and accurately reflect uncertainty.

The matching techniques that will be studied in this paper include deterministic record

linkage procedures developed by Ferrie (1996) and Abramitzky et al. (2012); abe; Abramitzky et al. (2019b), and Aizer et al. (2016), and multiple implementations of probabilistic record linkage, specifically the fastLink Enamorado et al. (2019), and machine learning approaches Feigenbaum (2016). Estimation techniques will include Anderson et al. (2019), Lahiri and Larsen (2005), and a fully Bayesian approach that I will develop in this paper.

The data used in this paper consist of the unmerged files from Aizer et al. (2016), which I will pre-process using the practices developed by Abramitzky et al. (2018). The parameter of interest is the average treatment effect of receiving a cash transfer on the long-term outcomes of children in poor families.

Section 2 describes the general problem that this paper seeks to address. Section 3 provides an overview of the matching techniques that I will study. Section 4 describes the estimation techniques. Section 5 will have simulations and results. Section 6 will have real data and results.

2 General problem

Suppose that the researcher would like to estimate θ_0 in a parametric model of the form

$$E[m(\mathbf{z_i}; \theta_0)] = 0 \tag{1}$$

where $m(\cdot)$ is a moment function and $\mathbf{z_i} = (x_i, y_i)$ is the data associated with an individual i sampled at random from the population of interest; however, instead of observing (x_i, y_i) pairs directly, the researcher observes two datasets. The first dataset D_1 contains variables x_i and identifiers w_i for individuals $i = 1, \ldots, N_1$. The second dataset D_2 contains outcomes y_j and identifiers w_j for individuals $j = 1, \ldots, N_2$.

To estimate (1) with standard econometric methods, the researcher must identify which

of the x_i and y_j refer to the same individuals. That is, she needs to recover the matching function $\varphi : \{1, \ldots, N_1\} \to \{1, \ldots, N_2\} \cup \varnothing$, where $\varphi(i) = j$ if individual i in dataset D_1 and individual j in dataset D_2 refer to the same entity; and $\varphi(i) = \varnothing$ if i does not have a match in D_2. If w_i and w_j identify individuals uniquely and do not change over time, then $\varphi(i) = j$ if and only if $w_i = w_j$; otherwise, $\varphi(i) = \varnothing$. However, if the identifiers are non-unique or prone to errors introduced by the record-generating process, then φ needs to be estimated, and inference about θ needs to be adjusted accordingly.

In statistics, the task of recovering φ is called *record linkage*. A record linkage procedure consists of a set of decisions about (i) selecting and standardizing the identifiers w_i and w_j, (ii) choosing which records to consider as potential matches (especially when $N_1 \times N_2 \times \dim(w_i)$ is large), (iii) defining what patterns of (w_i, w_j) constitute (partial) agreements, and (iv) designating (i, j) pairs as matches.[3] Each step of the record linkage process introduces the possibility that a true match is overlooked (Type II error), or that a false match is designated as correct (Type I error), and there is generally a tradeoff between reducing either one of the two (Abramitzky et al., 2019a; Doidge and Harron, 2018).

To fix ideas, suppose that θ_0 is the long-run impact of cash transfers on the longevity of children raised in poor families. The vector x_i includes family and child characteristics as observed in welfare program applications; and the outcomes y_j are constructed by calculating (day of death − day of birth) for all of the observations in a set of death records. For all i and j, the identifiers w_i and w_j include the individual's first and last name, and date of birth. Additionally, w_i includes i's place of birth; w_j includes j's place of death; and some w_i and w_j contain the individual's middle name or middle initial.

In this setting, an example of a (deterministic) record linkage procedure consists of:

[3]Note that this is the author's own definition. By contrast, Bailey et al. (2017) categorize historical linking algorithms (that match observations using name and age only) according to how they treat candidate pairs in the following four categories: (M1) a perfect, unique match in terms of name and age similarity; (M2) a single, similar match that is slightly different in terms of age, name, or both; (M3) many perfect matches, leading to problems with match disambiguation; (M4) multiple similar matches that are slightly different in terms of age, name or both.

(i) using a phonetic algorithm to standardize all string variables;

(ii) considering as potential matches only (i, j) pairs whose phonetically standardized names begin with the same letter, and whose birth years are within ± 2 years;

(iii) measuring agreements among names using Jaro-Winkler string distances, and weighing disagreements in birth year more than differences in birth month (and more than differences in birth day),

(iv) designating as matches all (i, j) pairs with scores calculated using the metrics in (iii) exceeding a pre-specified cutoff; and, if a record i has multiple possible matches j that exceed the cut-off, then choosing the match with the highest score (or picking a random match if there is a tie).

Another record linkage procedure could be defined using the same rules for steps (i)-(iii), but replace (iv) with a probabilistic matching rule that does not enforce one-to-one matching:

(iv*) use the Expectation-Maximization algorithm to compute "match weights" for each record pair; then designate as matches all pairs that exceed thresholds that are set to reflect specific tolerances for Type I and Type II error.

Except in rare cases, the matching function outputted by replacing (iv) with (iv*) will be different. Whereas the first procedure associates each x_i with at most one matched y_j, the second procedure may associate the same x_i with multiple y_j (in technical terms, this implies φ is a correspondence). The former case might use a standard GMM model to estimate θ; while the latter requires methods that associates multiple values of y_j with each x_i (Anderson et al., 2019). This example shows that not only do the estimates of θ likely depend on the estimates of φ, but also the *methods* for estimating θ may be also differ.

As observed by Bailey et al. (2017), record linkage procedures differ by the set of assumptions that motivate their use. However, all of the procedures discussed in this paper

will be studied under the following, common set of assumptions (with some departures later on):

1. (De-duplication) Within a given dataset, each observation refers to a distinct entity. That is, if two observations share the same identifier, they represent two different individuals.

2. (No unobserved sample selection) The observed x_i and y_j are random samples conditional on w_i and w_j, respectively. This means that all individuals with the same identifying information have equal probability of appearing in the sample.

3. There exists a unique $\theta_0 \in \Theta$ that satisfies the relationship in (1), that can be consistently estimated using standard econometric techniques if φ_0 is known.

The next section discusses in detail the exact record linkage techniques that will be studied, and their motivating assumptions.

3 Record Linkage Methods

Research on record linkage appears in many fields, such as statistics, computer science, operations research, and epidemiology, under many names, such as data linkage, entity resolution, instance identication, de-duplication, merge/purge processing, and reconciliation. As such, there are several books devoted to it study (Harron et al., 2015; Christen, 2012; Herzog et al., 2007), and dozens of commercial and open source systems software developed for its implementation (Köpcke and Rahm, 2010). Although insights from other fields have been slow to reach economics, recent working papers examine the impact of different record linkage techniques on inference using historic U.S. census data (Abramitzky et al., 2019a; Bailey et al., 2017); and the most recent version of the Handbook of Econometrics includes

a chapter on the "Econometrics of Data Combination" (Ridder and Moffitt, 2007).[4]

Record linkage techniques are broadly categorized as deterministic or probabilistic[5]; however, every deterministic linkage method has an equivalent probabilistic version (Doidge and Harron, 2018). For this reason, I will focus on developing estimation methods that use the probabilistic record linkage framework, but I will show how they can be applied to data that are linked deterministically.

3.1 Steps of the record linkage task

Here I will write a little bit more about each of the steps introduced in Section 2.

(i) selecting and standardizing the identifiers w_i and w_j

phonetic algorithms

(ii) choosing which records to consider as potential matches

blocking

(iii) defining what patterns of (w_i, w_j) constitute (partial) agreements

Jaro-winkler distances, for example

[4]Similar survey papers also exist in fields outside of economics, such as epidemiology and computer science (Doidge and Harron, 2018; Winkler, 1999). In fact, that record linkage is studied by many fields makes writing (and reading!) such surveys difficult, because authors are constantly writing the same things For example, Goldstein et al. (2012) prove similar results to those published in Hirukawa and Prokhorov (2018).

[5]define them here

3.1.1 (iv) designating (i, j) pairs as matches

this is where the differences between deterministic/probabilistic really arise!

3.2 Probabilistic Record Linkage

In describing the record linkage techniques implemented in this paper, I use notation from Fellegi and Sunter (1969). As before, consider two datafiles D_1 and D_2 that record information from two overlapping sets of individuals or entities. The files originate from two record-generating processes that may induce errors and missing values, so that identifying which individuals are represented in both D_1 and D_2 is nontrivial. I assume that individuals appear at most once in each datafile, so that the goal of record linkage is to identify which records in files D_1 and D_2 refer to the same entities.

Suppose that files D_1 and D_2 contain N_1 and N_2 records, respectively, and without loss of generality that $N_2 \geq N_1$. Denote also the number of entities represented in both files as N_M, so that $N_1 \geq N_M \geq 0$.

We say that the set of ordered record pairs $D_1 \times D_2$ is the union of two disjoint sets, *matches* (M) and *non-matches* (U):

$$M = \{(i, j) : i \in D_1, j \in D_2, i = j\}$$
$$U = \{(i, j) : i \in D_1, j \in D_2, i \neq j\}$$

Hence, the formal goal of record linkage is to identify whether an arbitrary record pair $(i, j) \in D_1 \times D_2$ belongs to M or U. Note that this task is identical to

To perform this task, each record pair is evaluated according to L different comparison criteria, which are the result of comparing data fields for records i and j. For example, if a record pair (i, j) represents two individuals, the pair may be evaluated according to whether

they share a first name or have the same birthday. These comparisons are represented by a *comparison vector*,

$$\boldsymbol{\gamma}_{\mathbf{ij}} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^\ell, \ldots, \gamma_{ij}^L)$$

where each comparison field γ_{ij}^ℓ may be binary-valued, as in "i and j have the same birthday," or use levels to account for partial agreement between strings (see **?**, for details). The models presented herein use only binary comparison vectors, however they may be extended to allow for partial agreement using the methods from **?**.

The probability of observing a particular configuration of $\boldsymbol{\gamma}_{\mathbf{ij}}$ can be modeled as arising from the mixture distribution:

$$P(\boldsymbol{\gamma}_{\mathbf{ij}}) = P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)p_M + P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)p_U \tag{2}$$

where $P(\gamma_{ij}|M)$ and $P(\gamma_{ij}|U)$ are the probabilities of observing the pattern γ_{ij} conditional on the record pair (i, j) belonging to M or U, respectively. The proportions p_M and $p_U = 1 - p_M$ are the marginal probabilities of observing a matched or unmatched pair. Applying Bayes' Rule, we obtain the probability of $(i, j) \in M$ conditional on observing $\boldsymbol{\gamma}_{\mathbf{ij}}$,

$$P(M|\boldsymbol{\gamma}_{\mathbf{ij}}) = \frac{p_M P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)}{P(\boldsymbol{\gamma}_{\mathbf{ij}})} \tag{3}$$

so that if we can estimate the variables in (2), we can estimate the probability that any two records refer to the same entity in (3).

As shown by Fellegi and Sunter (1969), it is possible to use the estimated probabilities to construct an "optimal" matching, given any threshold for false positive and false negative match rates. Conversely, the probabilities also allow us to estimate the false positive rate for any configuration of matches (**?**). Given the usefulness of the quantities in (3), the next sections will introduce two methods for estimating them.

3.3 Simplifying assumptions

Let $\mathbf{\Gamma} \equiv \{\boldsymbol{\gamma}_{\mathbf{ij}} : (i,j) \in X_1 \times X_2\}$ denote the set of comparison vectors for all records pairs $(i,j) \in X_1 \times X_2$. Note that $\mathbf{\Gamma}$ contains potentially $n_1 \times n_2$ elements, so that calculating $\mathbf{\Gamma}$ may be computationally expensive when X_1 or X_2 is large. In practice, researchers partition $X_1 \times X_2$ into "blocks," such that only records belonging to the same block are attempted to be linked, and records belonging to different blocks are assumed to be nonmatches. For example, postal codes and household membership are often used to define blocks when linking census files (**?**). Importantly, the blocking variables should be recorded without error, and sometimes there are none available. This paper assumes that no blocking is used; or, alternatively, that records are already divided into blocks that can be analyzed independently using the methods outlined below.

Conditional independence

In principle, we can model,

$$\boldsymbol{\gamma}_{\mathbf{ij}} \mid M \sim \text{Dirichlet}(\boldsymbol{\delta}_{\mathbf{M}})$$
$$\boldsymbol{\gamma}_{\mathbf{ij}} \mid U \sim \text{Dirichlet}(\boldsymbol{\delta}_{\mathbf{U}})$$

However, there are $2^L - 1$ possible configurations of each $\boldsymbol{\gamma}_{\mathbf{ij}}$, so that $\boldsymbol{\delta}_{\mathbf{M}}$ and $\boldsymbol{\delta}_{\mathbf{U}}$ may be very high-dimensional if we want to allow weights to vary across different comparison criteria.

A common assumption in the literature is that the comparison fields ℓ are defined so that γ_{ij}^{ℓ} are independent across ℓ conditional on match status. This implies:

$$P(\gamma_{ij}|C) = \prod_{\ell=1}^{L} P(\gamma_{ij}^{\ell}|C)^{\gamma_{ij}^{\ell}}(1 - Pr(\gamma_{ij}^{\ell}|C))^{1-\gamma_{ij}^{\ell}} \qquad C \in \{M, U\} \tag{4}$$

Hence the number of parameters used to describe each mixture class is reduced to L.

? have shown how to relax this assumption using log-linear models, but for now I assume conditional independence to ease computation.

3.4 Measuring record linkage performance

(could also summarize) They are evaluated according to matching rates, type I and type II error rates; robustness to selection/attrition?

List of record linkage methods Link of estimation methods

4 Estimation Methods

4.1 Standard assumptions

4.2 Relaxed assumptions

Matching variables correlated with variable of interest; true link not included

References

R. Abramitzky, L. Boustan, K. Eriksson, J. Feigenbaum, and S. Perez, "Automated linking of historical data," *NBER Working Paper*, 2019.

A. Aizer, S. Eli, J. Ferrie, and A. Lleras-Muney, "The long-run impact of cash transfers to poor families," *American Economic Review*, vol. 106, no. 4, pp. 935–71, April 2016. [Online]. Available: http://www.aeaweb.org/articles?id=10.1257/aer.20140529

R. Anderson, B. Honore, and A. Lleras-Muney, "Estimation and inference using imperfectly matched data," *Working paper*, August 2019. [Online]. Available: http://www.github.com/rachelsanderson/ImperfectMatching

H. Bleakley and J. Ferrie, "Shocking behavior: Random wealth in antebellum georgia and human capital across generations," *The Quarterly Journal of Economics*, vol. 131, no. 3, pp. 1455–1495, 2016. [Online]. Available: https://doi.org/10.1093/qje/qjw014

E. Nix and N. Qian, "The fluidity of race: Passing in the united states, 1880-1940," National Bureau of Economic Research, Working Paper 20828, January 2015. [Online]. Available: http://www.nber.org/papers/w20828

P. Lahiri and M. D. Larsen, "Regression analysis with linked data," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 222–230, 2005. [Online]. Available: http://www.jstor.org/stable/27590532

H. Goldstein, K. L. Harron, and A. M. Wade, "The analysis of record-linked data using multiple imputation with data value priors." *Statistics in medicine*, vol. 31 28, pp. 3481–93, 2012.

M. Bailey, C. Cole, M. Henderson, and C. Massey, "How well do automated linking methods perform? lessons from u.s. historical data," National Bureau of Economic Research, Working Paper 24019, November 2017. [Online]. Available: http://www.nber.org/papers/w24019

R. Abramitzky, R. Mill, and S. Perez, "Linking individuals across historical sources: a fully automated approach," National Bureau of Economic Research, Working Paper 24324, February 2018. [Online]. Available: http://www.nber.org/papers/w24324

K. Harron, A. Wade, R. Gilbert, B. Muller-Pebody, and H. Goldstein, "Evaluating bias due to data linkage error in electronic healthcare records," *BMC medical research methodology*, vol. 14, p. 36, 03 2014.

F. Scheuren and W. Winkler, "Regression analysis of data files that are computer matched - part ii," *Survey Methodology*, vol. 23, 01 1997.

I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, pp. 1183–1210, 1969.

J. P. Ferrie, "A new sample of males linked from the public use microdata sample of the 1850 u.s. federal census of population to the 1860 u.s. federal census manuscript schedules," *Historical Methods*, vol. 29, no. 4, pp. 141–156, 1 1996.

R. Abramitzky, L. P. Boustan, and K. Eriksson, "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration," *American Economic Review*, vol. 102, no. 5, pp. 1832–56, May 2012. [Online]. Available: http://www.aeaweb.org/articles?id=10.1257/aer.102.5.1832

R. Abramitzky, L. P. Boustan, and K. Eriksson, "To the new world and back again: Return migrants in the age of mass migration," *ILR Review*, vol. 72, no. 2, pp. 300–322, 2019. [Online]. Available: https://doi.org/10.1177/0019793917726981

T. Enamorado, B. Fifield, and K. Imai, "Using a probabilistic model to assist merging of large-scale administrative records," *American Political Science Review*, vol. 113, no. 2, p. 353?371, 2019.

J. J. Feigenbaum, "A machine learning approach to census record linking ?" 2016.

J. Doidge and K. Harron, "Demystifying probabilistic linkage," *International Journal for Population Data Science*, vol. 3, 01 2018.

K. Harron, H. Goldstein, and C. Dibben, *Methodological Developments in Data Linkage*. United States: John Wiley Sons Inc., 2015.

P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Publishing Company, Incorporated, 2012.

T. N. Herzog, F. J. Scheuren, and W. E. Winkler, *Data Quality and Record Linkage Techniques*, 1st ed. Springer Publishing Company, Incorporated, 2007.

H. Köpcke and E. Rahm, "Frameworks for entity matching: A comparison," *Data Knowledge Engineering*, vol. 69, pp. 197–210, 02 2010.

G. Ridder and R. Moffitt, "The Econometrics of Data Combination," in *Handbook of Econometrics*, ser. Handbook of Econometrics, J. Heckman and E. Leamer, Eds. Elsevier, January 2007, vol. 6, ch. 75. [Online]. Available: https://ideas.repec.org/h/eee/ecochp/6b-75.html

W. Winkler, "The state of record linkage and current research problems," *Statist. Med.*, vol. 14, 10 1999.

M. Hirukawa and A. Prokhorov, "Consistent estimation of linear regression models using matched data," *Journal of Econometrics*, vol. 203, no. 2, pp. 344 – 358, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304407617302464