Annotated bibliography

Rachel Anderson*

This Version: September 12, 2019

1 Challenges in historical record linkage

Historical U.S. data collected prior to the introduction of Social Security Numbers in 1935 often lack unique identifiers, so that identifying the set of individuals appearing in two or more datasets requires using characteristics such as name and reported age[1]. However, common names, coupled with poor data quality – caused by transcription and enumeration errors, age misreporting, mortality, under-enumeration, and migration between census years – makes linking records across two or more datasets with full certainty impossible. (**?**).

For example, while matching children listed on mother's welfare program applications with Social Security Administration death records recorded decades later, **?** find X eprcent of indivdiauls without a match, and X cases wehre multiple death records seem to refer to the same individual. Similarly, Goeken et al. (2017) document that in two enumerations of St. Louis in the 1880 Census, nearly 46 percent of first names are not exact matches, and the Early Indicators project notes that 11.5 percent of individuals in the Oldest Old sample have a shorter first name in pension records than in the original Civil War enlistment records

*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.

(Costa et al. 2017).

If the researcher is willing to assume that errors produced by the record generating process are uncorrelated with the variable of interest, then standard record linkage techniques perform fine, so long as they account for uncertainty from the matching process. More challenging is the case where errors in the record generating process are correlated with the variables of interest. This is exactly the case in **?** by low literacy rates and regional variations in names, as well as the record digitization process itself which introduces yet another possible source of error.

To illustrate this point, consider an illiterate individual from Louisiana with the surname of Thibideaux, who chooses to move to another state, would likely have his name spelled phonetically as Tibido. Researchers use a variety of techniques – phonetic algorithms, string comparators, and probabilistic record linkage – to link records from different files despite errors that arise in the record generating process.

Compared to other fields where record linkage is a research goal in itself, record linkage in economics is seen as prerequisite to answering economic questions. Economists contribute to the record linkage literature by focusing on historic data; with care for the impact it has on subsequent inference.

This is problem bc Neter, Maynes, and Ramanathan (1965): small mismatch errors in finite population sampling can lead to a substantial bias in estimating the relationship between response errors and true values

Matching methods and estimation methods are studied separately, but really ought to be studied together.

2 Record Linkage Methods

Record linkage is such a common yet difficult task that there are several books devoted to its study (**???**), and dozens of commercial and open source systems software developed for its implementation (**?**). In economics, a recent series of working papers examine the effects of different record linkage techniques on subsequent inference (**??**). However, similar survey papers also exist in fields outside of economics, such as epidemiology and computer science (**??**). 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.[2]

Disciplinary differences aside, record linkage methods can be broadly divided into deterministic and probabilistic techniques. That said, the choice between using deterministic and probabilistic methods is a false dichotomy, since for every deterministic linkage method there is an equivalent probabilistic one (**?**). The main differences among record linkage methods differ in the choices they make about how to represent the data, which comparisons to consider as possible links, and whether to enforce one-to-one matching or allow multiple matches. Common among all techniques is the necessity of (i) setting thresholds, (ii), blocking data, and (iii) . See existing paper.

The father of modern record linkage, Winkler in **?** says: "Although individuals have introduced alternative classification methods based on Support Vector Machines, decision trees and other methods from machine learning, no method has consistently outperformed methods based on the Fellegi?Sunter model, particularly with large day-to-day applications with tens of millions of records"

? review literature on historical record linkage in US and examines performance of automated record linkage algorithms with two high-quality historical datasets and one synthetic ground truth. They conclude that no method consistently produces representative samples;

[2] **?** prove similar results to those published in **?**.

machine linking has high number of false links and may introduce bias into analyses.

? have guide for researchers in the choice of which variables to use for linking, how to estimate probabilities, and then choose which records to use in the analysis. Created R code and stata command to implement the method

? evaluate different automated methods for record linkage, specifically deterministic (like Ferrie and ABE papers), machine learning Feigenbaum approach, and the AMP approach with the EM algorithm. Document a frontier between type I and type II errors; cost of low false positive rates comes at cost of designating relatively fewer (true) matches. Humans typically match more at a cost of more false positives. They study how different linking methods affect inference – sensitivity of regression estimates to the choice of linking algorithm. They find that the parameter estimates are stable across linking methods. Find effect of matching algorithm on inference is small.

2.1 Matching Methods

? 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

Historical linking algorithms generally treat M1 cases as matches, but differ in how they treat M2, M3, and M4 candidate pairs. Generally, differences in M2 are solved deterministically by setting fixed-year band tolerances for matches, and probabilistically by estimating weights for the relative importance of age vs. name agreement. Multiple matches in M3/M4 are ignored,

picked at random, given equal weights, or given weights proportional to the probability of being the true match. Table X below provides an overview of methods in literature based on these dimensions.

[Insert table here] Table includes

Talk also about how to evaluate these matching methods – what is desirable? How to estimate error rates ex post!

- Ferrie 1996, Abramitzky, BOustan and Eriksson (2012 2014 2017) Deterministic. Conservative methods require no other potential match with same name within a 5-year band , Nix and Qian
- Semi-automated Feigenbaum, Ruggles et al

3 Estimation Methods

3.1 ??

Scheuren and Winkler (1993) give a bias adjustment – use linkage probabilities to correct for failures in the matching step. However, **?** (1997) is the first paper to propose a unified approach between the linkage and the analysis. But no theoretical results. lacks a convergence proof, etc.

3.2 ?

? take as input two files are linked by a computerized record linkage technique (CRL). The true data pairs (x_i, y_i) are not observable; instead, the CRL produces pairs (x_i, z_i) in

which z_i may or may not correspond to y_i. The (true) regression model is:

$$y_i = x_i'\beta + \epsilon_i,\ E[\epsilon_i] = 0,$$
$$\text{var}(\epsilon_i) = \sigma^2,\ \text{cov}(\epsilon_i\epsilon_j) = 0$$

but the researcher estimates this model with z_i as the dependent variable, where

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \text{ for } j \neq i,\ j = 1, \ldots, n \end{cases}$$

and $\sum_{j=1}^{n} q_{ij} = 1,\ i = 1, \ldots, n$. Define $\mathbf{q_i} = (q_{i1}, \ldots, q_{in})'$. The naive least squares estimator of β, which ignores mismatch errors, is given by:

$$\hat{\boldsymbol{\beta}}_N = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$$

An alternative to this naive estimator is one that minimizes the sum of absolute deviations, which decrease the influence of outliers and hence should decrease the impact of erroneously paired predictor and response values.

Note that $E(z_i) = \mathbf{w_i'}\boldsymbol{\beta}$, with $\mathbf{w_i} = \mathbf{q_i'}\mathbf{X_i} = \sum_{j=1}^{n} q_{ij}x_j'$, and so the bias of $\hat{\boldsymbol{\beta}}_N$ is given by

$$\text{bias}(\hat{\beta}_N) = [(X'X)^{-1}X'QX - I]\beta$$

Hence, if $Q = I$, then $\hat{\beta}_N$ is unbiased. This is equivalent to giving all potential matches the same weight (i.e. treating all matches as equally likely to be correct), as discussed in **?**.

In order to reduce the bias of $\hat{\beta}_N$, Scheuren and Winkler (1993) observed that

$$\text{bias}(\hat{\beta}_N|y) = E[(\hat{\beta}_N - \beta)|y] = (X'X)^{-1}X'B,$$

where $B = (B_1, \ldots, B_n)'$ and $B_i = (q_{ii} - 1)y_i + \sum_{j \neq i} q_{ij} y_j = \mathbf{q_i'y} - y_i$, which is the difference between a weighted average of responses from all observations and the actual response y_i. Thus, if an estimator $\hat{B}$ is available, the SW estimator is given by:

$$\hat{\beta}_{SW} = \hat{\beta}_N - (X'X)^{-1}X'\hat{B}$$

SW give a truncated estimator of B_i using the first and second highest elements of the vector q_i, $\hat{B}_i^{TR} = (q_{ij_1} - 1)z_{j_1} + q_{ij_2}z_{j_2}$, that can also be written more generally for an arbitrary number of elements of q_i. This means that $\hat{\beta}_{SW}$ is not unbiased, but, if the probability is high that the best candidate link is the true link, then the truncation might produce a very small bias.

Using $E(z) = W\beta$, **?** propose an exactly unbiased estimator of β:

$$\hat{\beta}_U = (W'W)^{-1}W'z$$

where $W = (w_1, \ldots, w_N)'$, $w_i = q_i'X_i$ as above. They suggest using a truncated version of w_i, $w_i^{TR} = q_{ij_1}x_{j_1} + q_{ij_2}x_{j_2}$. **?** use estimates of Q obtained from applying the Fellegi-Sunter/EM procedure, and observe that replacing Q with $\hat{Q}$ yields unbiased estimates of β whenever $\hat{Q}$ can be assumed to be independent of z. They argue that this is expected to be true in most applications, because the distribution of matching variables (e.g. first and last name, age), which determines the distribution of $\hat{Q}$, is usually independent of the response variable y (e.g. income), and hence of z.

Importantly, this assumption does not hold in some economics applications, such as the racial "passing" example from **?**.

? conclude that in simulations, least median regression is not sufficient to guard against matching errors, whereas the method of SW (1003) made a useful adjustment. Their method performed well across a range of situations,a nd the bootstrap procedure is useful for reflect-

ing uncertainty due to matching.

3.3 Prior-Informed Imputation (?)

?, among others (**?**), show that choosing the match with the highest probabilistic weight leads to biased estimates. This bias increases as the threshold for acceptance increases, and when there is an association between the linkage error probabilities and the variables in the model of interest. An example of this is Lariscy (YEAR).

Prior-informed imputation (PII), proposed in **?**, aims to select the correct value for variables of interest, rather than accepting a single complete record as a link. Information from match probabilities in candidate linking records (the prior) is combined with information in unequivocally linked records.

An example of PII in practice is **?**.

NOTABLY, uses 1-1 matching (no multiple matches). Also assume that the probability of matching on MV values is independent of the variable of interest. They recognize this is not often the case; i.e. different hospitals may have different distributions of the variable of interest, quality of recorded identifying info may also different by center, creating lack of independence between probability of a match and variable of interest. This is shown to be a problem in **?**

Use the usual FS/EM procedure to estimate probabilities of candidate pairs referring to a match. The result is each record i is associated with $\{y_{ij}\}$ and probabilities p_{ij}. Assume, wlog, that all variables follow a joint multivariate normal distribution[3] A lower threshold can be chosen so that records with probabilities less than a threshold are ignored. In practice, they recommend ignoring records that have no match on any matching variable; regard these

[3]If this is not the case for some of the observed variables, then a joint MVN distribution can be obtained using the latent MVN trick (for categorical variables, imputed values are back-transformed to their original scales.

records as having missing variables and use standard MI.

Denote distribution of variables in linking datafile A (y in ahl framework), conditional on variables in the file of interest B that they are linking to, by $f(Y^{A|B})$. Conditioning is on responses and any covariates in the imputation model, includes variables from A that are treated as auxiliary predictor variables in the imputation model. This conditional distribution is also multivariate normal.

For each record i, we compute a modified prior probability π_{ij} which is the likelihood component $f(Y^{A|B})$ multiplied by the prior p_{ij}, so that $\pi_{ij} \propto f(Y^{A|B})p_{ij}$. The normalized set π_i comprises the modified probability distribution for each i record in A.

Set a lower threshold for accepting a record as a true link, and if any records exceeds this, we choose that with the largest probability. If no record exceeds the threshold, then we regard the data values as missing and use standard multiple imputation. "The largest gain can be expected to arise when the probability of a link is associated with the values of the variables to be transferred. When the MAR assumption discussed earlier holds, then given a high enough threshold, the proposed procedure will produce unbiased estimates" Incorporating the likelihood component can be expected to lead more often to the correct record exceeding a given threshold.

Proposed method for FIXING THIS ASSUMPTION! Thus far assumed true matching record is located within the B file. But this may not be the case. Assume we have an estimate of mortality rate of individuals in A, π_d. If a proportion of the A file $\pi_m < \pi_d$ are unequivocally matched, then the probability that a randomly chosen remaining record in B is not a death from the A file is $\pi_r = 1 - (\pi_d - \pi_m)$. Therefore multiply p_i by $1 - \pi_r$ and add an extra pseudo-record with probability π_r with an associated code for a surviving patient.

3.4 ?

This is a perfect application for estimating the mean using our methods! That is the goal of **?**.

? study racial passing by linking individual U.S. census records to determine whether an individual's recorded race changed from one census to the next. To achieve higher match rates than those of previous studies[4], the authors develop methods for including individuals with multiple potential matches. These methods include selecting one match at random, and selecting the match that produces an upper/lower bound for estimating the "passing" rate.

? also use the unmatched individuals from their data to calculate absolute bounds for the population passing rates. For a given algorithm, the absolute upper bound is obtained by using the "upper bound" configuration of data, combined with assuming that all unmatched individuals passed. The lower bound is obtained in the same way, assuming that none of the excluded individuals passed.

? argue that increasing the match rate improves the bounds around any true population statistic, even though their methods introduce random measurement error in the estimand. Below is a visual of their complex blocking strategy.

3.5 ?

? assign equal probability of winning (matched variable equal to $1/n$) to all n individuals matched to the same winner. The goal is to estimate the treatment effect of winning a parcel in the lottery by comparing mean outcomes for winners and losers in a simple bivariate regression with a dummy varialbe for winning a parcel on the right-hand side. Here, winning

[4]The authors match 61-67 percent of individuals. ABE (2012), Hornbeck and Naidu (2014), Long and Ferrie (2013), Mill and Stein (2012) have match rates around 30, 24, 22, 11-34 percent respectively

Figure 1: ? blocking strategy

Figure A.8: Average Number of Potential Matches

the lottery is coded as 0 or $1/n$, where n is the number of matches for person i. [Think about how this compares to ahl method]

3.6 Other ideas to include?

? give OLS methods for nearest neighbor matched samples. Not explicitly the probabilistic record linkage framework, although the authors use this term. Could apply their methods for an appropriately defined metric for matching variables (i.e. Jaro-Winkler string distance), and use a nearest neighbor matching rule to assign 1-1 matches (although some y may get mapped to multiple x.

Enamorado procedures Survey paper from handbook of econometrics

"However, the analytic estimates of precision in Lahiri and Larsen (2005) are poor for 1-1 probabilistic linkage (Chipperfield and Chambers 2015)" As a quality measure, Christen (2012) suggests precision, which is the proportion of links that are true matches. Winglee et al. (2005) use a simulation-based approach, Simrate to estimate linkage quality. Their method uses the observed distribution of data in matched and non-matched pairs to generate a large simulated set of record

FROM https://arxiv.org/pdf/1901.04779.pdf

Simulation Idea

Could use only simulation data, with variety of possible biases, motivated by the applications above. For example, motivated by N-Q, introduce error with geographical relocation. Then test which techniques are robust to these types of sample selection/error.

I will compare estimates of Type I/ Type II error to ACTUAL Type I/ Type II error

rates, and say that authors need to estimate their errors!! Not just report the match rate. Use estimates from Chipperfield (2018) maybe https://onlinelibrary.wiley.com/doi/epdf/10.1111/insr.12246

4 Bird's eye view of literature

Record linkage surveys

- Books on record linkage: **???**
- Review of free software tools: **?**
- Survey of RL techniques in economics : **?**, **?**
- Surveys by the"father" and "mother" of modern record linkage (**?**)
- Comparison of deterministic vs. probabilistic matching methods (**?**)

Matching methods

- Probabilistic Record Linkage (**??**)
- Prior-informed imputation (**?**)
- Bayesian "Beta" Record Linkage (**?**)
- Deterministic methods for economics (**???**)
- Deterministic methods for bounding parameter (**?**)
- Allow multiple matches by imputing equal weight values (**?**)

Estimation methods

- Integrated matching and estimation procedures
  - Iterative matching, regression analysis, imputation (**?**)
  - Bayesian matching and population size (**?**)
- Bias corrections for regression with matched data
  - Bias correction for regression (**??**)
  - Weighted least squares-based bias correction for linear and logistic regression (**?**)
  - Bias correction for NN-matching (**?**)
- GMM with multiple matches (**?**)