Annotated bibliography

Rachel Anderson*

This Version: September 4, 2019

1 Challenges in matching historical datasets

Record linkage in economics is most often studied by those who use historical U.S. data prior to the introduction of social security numbers (Abramitzky et al., 2019; Aizer et al., 2016; Ferrie, 1996). These data tend to lack unique identifiers, so that individuals are primarily identified by their first and last name, and age[1], which are rarely unique within a given population. The non-uniqueness problem is exacerbated by typographical error, caused 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.

For example, Nix and Qian (2015) remark that 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. 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

*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]Sometimes middle initial, birthplace, parent's name and birthplace, exact birthdate etc. are available; but data quality vary by application.

records (Costa et al. 2017).

Researchers use phonetic algorithms to account for differences in spelling.

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

2 Relevant Survey Papers

Record linkage is such an intricate and frequent challenge that there are several books devoted to its study, and dozens of commercial and freely available software devoted to the task (Christen, 2012; Herzog et al., 2007). The research and open source systems include those developed by census bureaus, computer scientists, epidemiologists, statisticians, political scientists, and – only more recently – economists. A recent survey by Köpcke and Rahm (2010) evaluates eleven such systems.

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.

Christen (2012) 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"

Bailey et al. (2017) 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; machine linking has high number of false links and may introduce bias into analyses.

Abramitzky et al. (2018) 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

Abramitzky et al. (2019) 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.

3 Matching Methods

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

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

4 Estimation Methods

4.1 Lahiri and Larsen (2005)

Lahiri and Larsen (2005) take as input two files are linked by a computerized record linkage technique. 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 Anderson et al. (2019).

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 $B_i = (q_{ii} - 1)y_i + \sum_{j\neq i} q_{ij}y_j = \mathbf{q_i'}\mathbf{y} - y_i$, 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 $E(z) = W\beta$, Lahiri and Larsen (2005) propose an exactly unbiased estimator of β:

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

that can be obtained by using a truncated version of w_i, $w_i^{TR} = q_{ij_1}x_{j_1} + q_{ij_2}x_{j_2}$. Lahiri and Larsen (2005) 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 Nix and Qian (2015).

Lahiri and Larsen (2005) 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 reflecting uncertainty due to matching.

1. Anderson et al. (2019) use all matches with weight 1/n
2. Nix and Qian (2015) pick one match at random, then upper/lower bounds
3. Bleakley and Ferrie (2016) weight treatment variable as 1/n
4. exploratory methods (if time)

Scheuren and Winkler (1993): propose method for adjusting for bias of mismatch error in OLS SW (1997, 1991): iterative procedure that modifies regression and matching results for apparent outliers 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

Important Applications

Nix and Qian (2015) 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[2], 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.

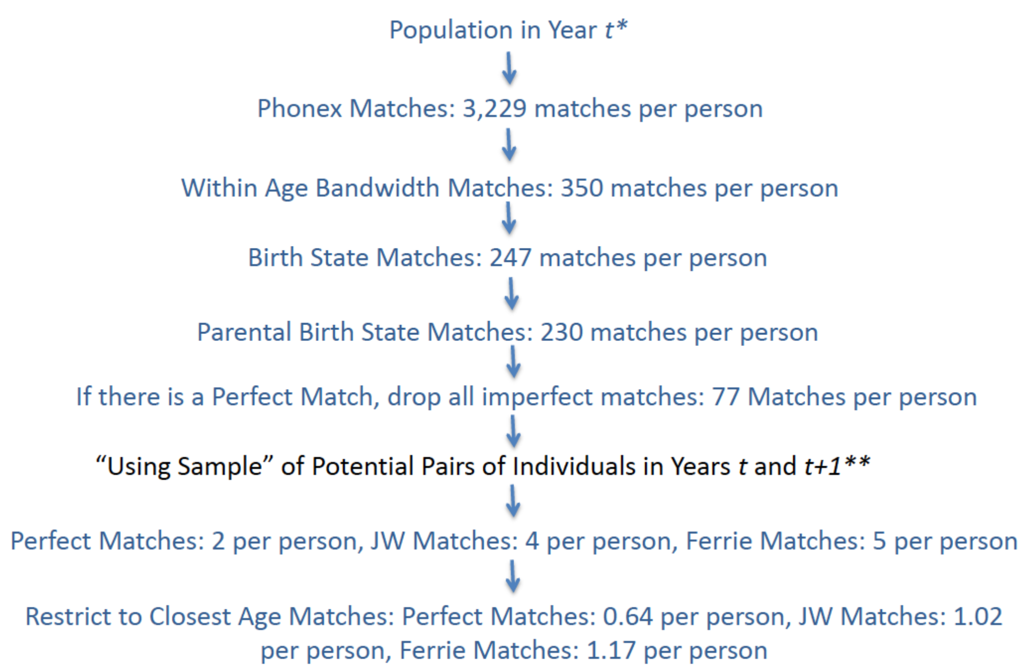
[2] The authors match 61-67 percent of individuals. ABE (2012), Hornbeck and Naidu (2014), Long and Ferrie (2013), Mill and Stein (2012) have match rates around 30, 24, 22, 11-34 percent respectively

Nix and Qian (2015) 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.

Nix and Qian (2015) 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.

Figure 1: Nix and Qian (2015) blocking strategy

Figure A.8: Average Number of Potential Matches



Bleakley and Ferrie (2016) 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 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]

Matching Method	Type	Blocking Variables	Matching Variables	String Distance	Matching Rule
Ferrie (1996), Abramitzky, Boustan and Eriksson (2012, 2014, 2017)	Deterministic	Year of birth	Any time-invariant characteristic	NYSIIS, Jaro-Winkler	Iteratively search for unique matches, increasing the tolerance for error with each iteration. Accept matches that are sufficiently unique.
Aizer et al. (2016)	Deterministic	None	First/last name, middle initial, day, month, and year of birth	SOUNDEX, SPEDIS	Designate as a possible match any observation pair that satisfied pre-specified criteria. Group matches based on quality, and retain only "best" matches (can be multiple)
Abramitzky, Mill, and Pérez (2019)	Probabilistic				
Feigenbaum (2016)	Machine Learning				
Nix and Qian (2015)	Deterministic	Phonex-standardized name, age, birth state, parental birth states	Name, age, race	SOUNDEX, Jaro-Winkler	3 matching algorithms $\times$ 4 sample restrictions produce 12 samples used in analysis. analysis. Multiple matches handled by selecting one potential match at random, or choosing the match that induces an"upper"/"lower" bound for object of interest

Figure 2: default

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

References

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

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.

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.

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

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

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

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

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