Annotated bibliography

Rachel Anderson[*]

This Version: September 4, 2019

Historical Matching Challenges

Primary variable for matching is an individual's name, which are rarely unique within a given population. Low literacy rates mean names are often misspelled. Misspelling also arises due to geographical relocation and regional variation in names. "For example, 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." (Nix and Qian 2015). Researchers use phonetic algorithms to account for these differences

For example, 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 (Costa et al. 2017).

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

(Meta) MatchingPapers

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.

? 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.

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

Estimation Methods

1. Lahiri and Larsen (2005) adjustment for regression using estimated probabilities from E-M algorithm (can use ARP implementation)

2. Anderson et al. (2019) use all matches with weight 1/n

3. Nix and Qian (2015) pick one match at random, then upper/lower bounds

4. Bleakley and Ferrie (2016) weight treatment variable as 1/n

5. 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[1], 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.

---

[1]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

Population in Year t*

↓

Phonex Matches: 3,229 matches per person

↓

Within Age Bandwidth Matches: 350 matches per person

↓

Birth State Matches: 247 matches per person

↓

Parental Birth State Matches: 230 matches per person

↓

If there is a Perfect Match, drop all imperfect matches: 77 Matches per person

↓

"Using Sample" of Potential Pairs of Individuals in Years t and $t+1$**

↓

Perfect Matches: 2 per person, JW Matches: 4 per person, Ferrie Matches: 5 per person

↓

Restrict to Closest Age Matches: Perfect Matches: 0.64 per person, JW Matches: 1.02 per person, Ferrie Matches: 1.17 per person

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 × 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

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. 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