Population Research with Linked Data: Guide to Inference

Methods and Analysis of Linked Data | PAA 2025

Casey F. Breen[1] Won-tak Joo[2]

April 24, 2025

[1]University of Oxford
[2]University of Florida

The growth of linked data in the social sciences

- Explosion in publicly-available linked census and admin data resources (Ruggles et al., 2020; Genadek and Alexander, 2022; Goldstein et al., 2021; Abramitzky et al., 2020)

The growth of linked data in the social sciences

- Explosion in publicly-available linked census and admin data resources (Ruggles et al., 2020; Genadek and Alexander, 2022; Goldstein et al., 2021; Abramitzky et al., 2020)
  - Much lower barriers to entry (500+ social science papers per year)

The growth of linked data in the social sciences

- Explosion in publicly-available linked census and admin data resources (Ruggles et al., 2020; Genadek and Alexander, 2022; Goldstein et al., 2021; Abramitzky et al., 2020)
  - Much lower barriers to entry (500+ social science papers per year)
- Large and important body of methodological research on improving record linkage (Ruggles, Fitch and Roberts, 2018; Bailey et al., 2020; Hwang and Squires, 2024; Postel, 2023; Abramitzky et al., 2020; Helgertz et al., 2022)

Less Methodological Attention to Inference

- Some guidance exists for inference with linked data (Bailey, Cole and Massey, 2019; Bailey et al., 2020)

Less Methodological Attention to Inference

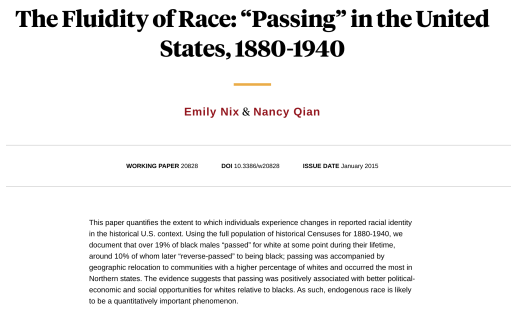
- Some guidance exists for inference with linked data (Bailey, Cole and Massey, 2019; Bailey et al., 2020)
- No formal framework or consensus on best practices for inference under linkage error

Less Methodological Attention to Inference

- Some guidance exists for inference with linked data (Bailey, Cole and Massey, 2019; Bailey et al., 2020)
- No formal framework or consensus on best practices for inference under linkage error
- This study introduces a **framework** to unpack bias introduced by false and missed matches

Less Methodological Attention to Inference

- Some guidance exists for inference with linked data (Bailey, Cole and Massey, 2019; Bailey et al., 2020)
- No formal framework or consensus on best practices for inference under linkage error
- This study introduces a **framework** to unpack bias introduced by false and missed matches



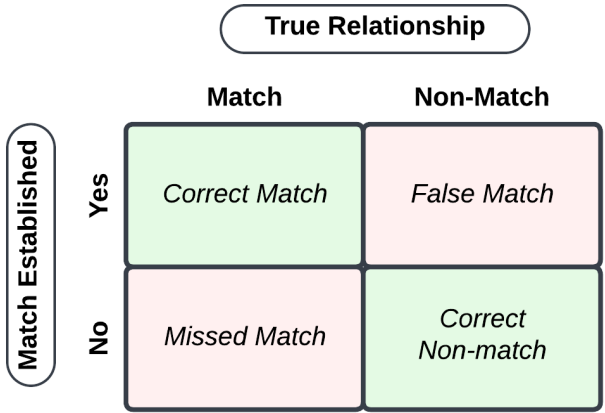
The Fluidity of Race: "Passing" in the United States, 1880-1940

Emily Nix & Nancy Qian

WORKING PAPER 20828 **DOI** 10.3386/w20828 **ISSUE DATE** January 2015

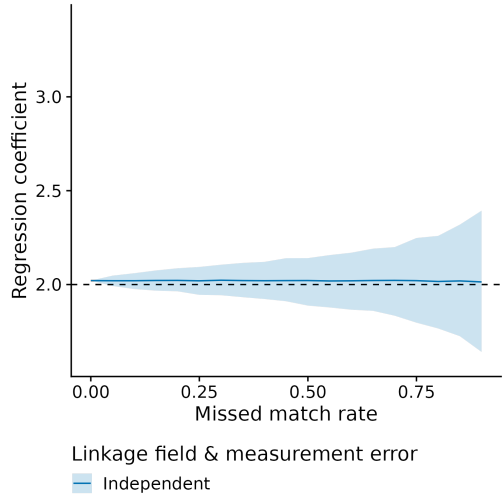
This paper quantifies the extent to which individuals experience changes in reported racial identity in the historical U.S. context. Using the full population of historical Censuses for 1880-1940, we document that over 19% of black males "passed" for white at some point during their lifetime, around 10% of whom later "reverse-passed" to being black; passing was accompanied by geographic relocation to communities with a higher percentage of whites and occurred the most in Northern states. The evidence suggests that passing was positively associated with better political-economic and social opportunities for whites relative to blacks. As such, endogenous race is likely to be a quantitatively important phenomenon.

Two types of linkage errors



Missed Matches

- Smaller sample size $\rightarrow$ reduced statistical power and larger uncertainty
- Potential **selection bias** in records that are successfully linked



Conceptual parallel with non-probability sampling

In non-probability sampling, from a population U:

$$\pi_i = P(i \in S | i \in U) \tag{1}$$

where

- S is the sample

Conceptual parallel with non-probability sampling

In non-probability sampling, from a population U:

$$\pi_i = P(i \in S | i \in U) \tag{1}$$

where

- S is the sample
- π_i is inclusion probability in the sample

Conceptual parallel with non-probability sampling

- Unknown π_i complicates population parameter estimation and inference
- Analogous to bias from linkage errors in linked data analysis
- Pick correct reference population for weighting...

Non-Probability Toolkit

- Post-stratification weighting
- Raking
- Inverse probability weighting*
- Various matching approaches...

False matches - descriptive rates

Ideal case (no false matches):

$$R = \frac{O}{N} \tag{2}$$

- R: Observed rate (e.g., event rate)
- O: Number of observed outcomes/events
- N: Sample size (denominator)

False matches - descriptive rates

$$R' = \underbrace{R_{\text{true}} \times (1 - f_r)}_{\text{Contribution of True Matches}} + \underbrace{R_{\text{false}} \times f_r}_{\text{Contribution of False Matches}} \tag{3}$$

- R_{true}: Rate for true matches
- R_{false}: Rate for false matches
- f_r: False match rate

False matches — regression coefficients

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{4}$$

where:

$$\hat{\beta}_1 = \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(X)} \tag{5}$$

False matches — regression coefficients

$$\hat{\beta'}_1 = \frac{(1 - f_r)(\mathsf{Cov}(X, Y)) + (f_r)\,(\mathsf{Cov}(X_{\mathsf{false}}, Y_{\mathsf{false}}))}{\mathsf{Var}(X)} \tag{6}$$

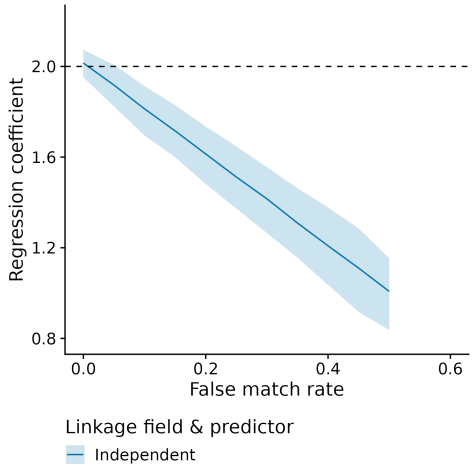
Regression framework

$$\hat{\beta}_1' = \frac{(1 - f_r) \cdot \mathsf{Cov}(X, Y) + f_r \cdot \cancel{\mathsf{Cov}(X_{\mathsf{false}}, Y_{\mathsf{false}})}}{\mathsf{Var}(X)} \tag{7}$$

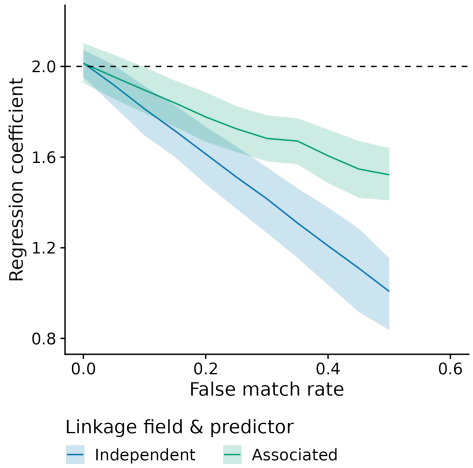
$$= \frac{(1 - f_r) \cdot \mathsf{Cov}(X, Y)}{\mathsf{Var}(X)} \tag{8}$$

$$= \hat{\beta}_1(1 - f_r) \tag{9}$$

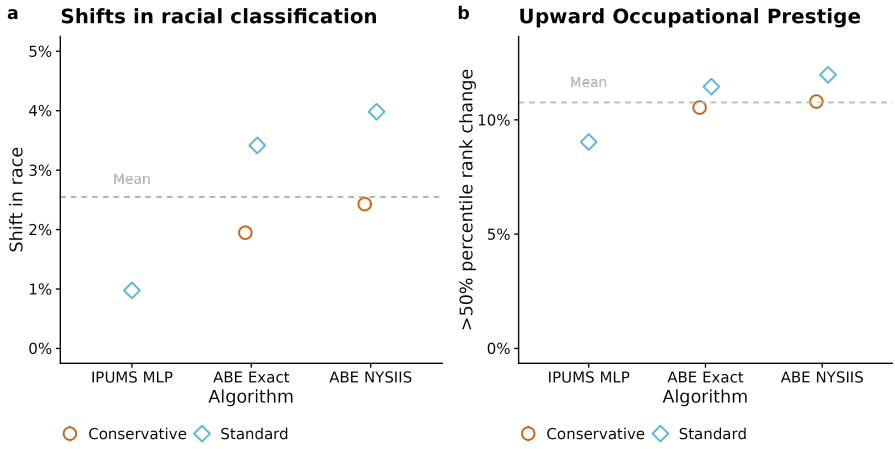
Illustrative simulation



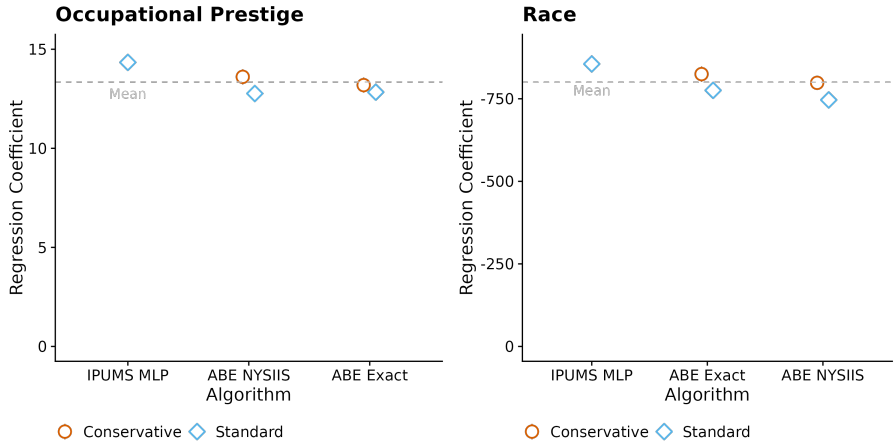
Illustrative simulation



Empirical Results



Empirical Results — regression on wage/salary income



How Do We Practically Address False Matches?

- **Validation variable:**
  - Variable not used in the linkage process but available in both datasets, such as middle initial, month of birth (Bailey, Cole and Massey, 2019)
  - Disagreement suggests false match...
- **Sensitivity analysis:**
  - Vary assumed false match rate f_r
  - Re-estimate key coefficients under plausible error scenarios

Case Study — Racial Passing by Birth Cohort

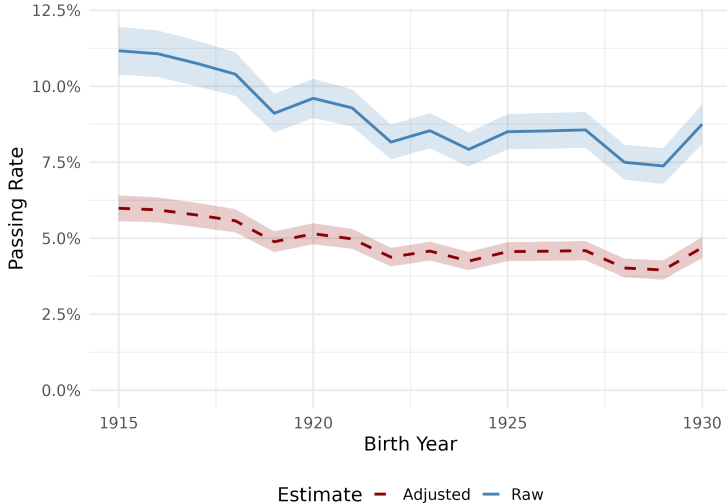
Data:

- Link individuals from the 1940 Census to the Social Security Numident file (CenSoc-Numident)

Validation and Adjustment Steps:

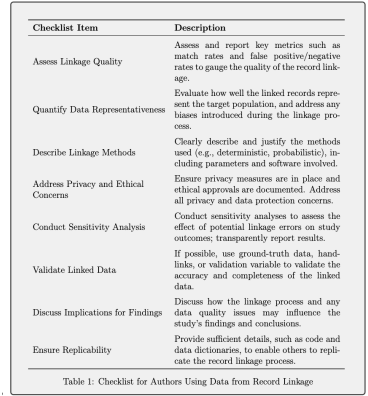
1. Identify cases with a middle initial in both datasets (25% of sample)
2. Use middle initial agreement to estimate the false match rate
3. Compute an adjustment factor based on this validation subsample
4. Apply the adjustment factor to correct estimates in the full linked sample

Empirical results — rates of racial passing



Reporting standards - Checklist for linked data

- Checklist for promoting transparency and replicability in record linkage science
- Key items
  1. Describe linkage method
  2. Quantify data representativeness
  3. Discuss implications of linkage errors for findings



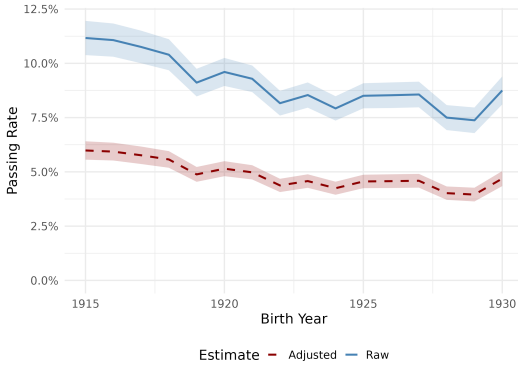
Checklist Item	Description
Assess Linkage Quality	Assess and report key metrics such as match rates and false positive/negative rates to gauge the quality of the record linkage.
Quantify Data Representativeness	Evaluate how well the linked records represent the target population, and address any biases introduced during the linkage process.
Describe Linkage Methods	Clearly describe and justify the methods used (e.g., deterministic, probabilistic), including parameters and software involved.
Address Privacy and Ethical Concerns	Ensure privacy measures are in place and ethical approvals are documented. Address all privacy and data protection concerns.
Conduct Sensitivity Analysis	Conduct sensitivity analyses to assess the effect of potential linkage errors on study outcomes; transparently report results.
Validate Linked Data	If possible, use ground-truth data, hand-links, or validation variable to validate the accuracy and completeness of the linked data.
Discuss Implications for Findings	Discuss how the linkage process and any data quality issues may influence the study's findings and conclusions.
Ensure Replicability	Provide sufficient details, such as code and data dictionaries, to enable others to replicate the record linkage process.

Table 1: Checklist for Authors Using Data from Record Linkage

Conclusion

- **Framework** for unpacking errors in inference with linked data:
  - Missed matches can may introduce selection bias—but can apply full non-probability toolkit
  - False matches are more challenging to account for
  - We can estimate the bias they introduce if we know the (1) false match rate and (2) covariance / association among false matches
- **Record linkage checklist**: a checklist for social science research with linked data

Questions?



caseyfbreen

casey.breen@demography.ox.ac.uk

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Santiago Pérez and Myera Rashid. 2020. "Census Linking Project: Version 1.0.".

Bailey, Martha, Connor Cole and Catherine Massey. 2019. "Simple Strategies for Improving Inference with Linked Data: A Case Study of the 1850–1930 IPUMS Linked Representative Historical Samples." *Historical methods* 53(2):80.

Bailey, Martha, Connor Cole, Morgan Henderson and Catherine Massey. 2020. "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data." *Journal of economic literature* 58(4):997–1044.

Genadek, Katie R. and J. Trent Alexander. 2022. "The Missing Link: Data Capture Technology and the Making of a Longitudinal U.S. Census Infrastructure." *IEEE Annals of the History of Computing* pp. 1–10.

Goldstein, J. R., M. Alexander, C. Breen, A. Miranda González, F. Menares, M. Osborne, M. Snyder and U. Yildirim. 2021. "Censoc Project." *CenSoc Mortality File: Version 2.0. Berkeley: University of California* .

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J Thompson, Steven Ruggles and Catherine A. Fitch. 2022. "A New Strategy for Linking U.S. Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55(1):12–29.

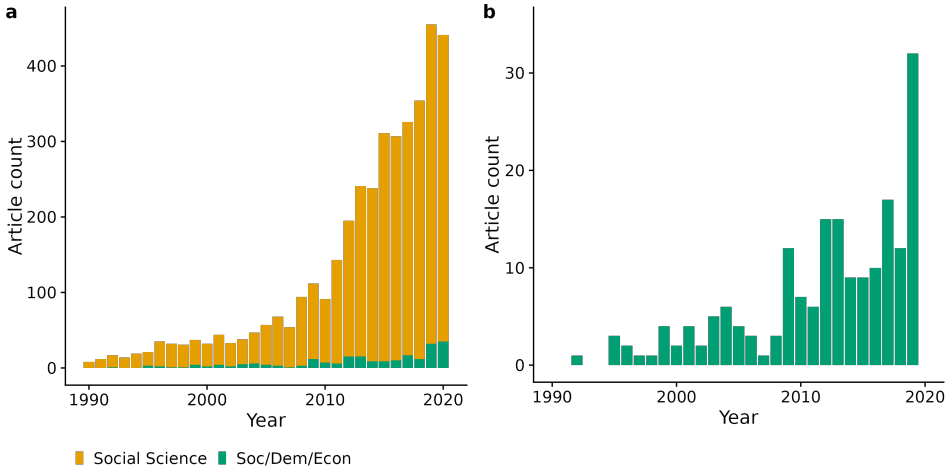
Hwang, Sam Il Myoung and Munir Squires. 2024. "Linked Samples and Measurement Error in Historical US Census Data." *Explorations in Economic History* 93:101579.

Postel, Hannah M. 2023. "Record Linkage for Character-Based Surnames: Evidence from Chinese Exclusion." *Explorations in Economic History* 87:101493.

Ruggles, Steven, Catherine A. Fitch and Evan Roberts. 2018. "Historical Census Record Linkage." *Annual Review of Sociology* 44(1):19–37.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Mathew Sobek. 2020. "IPUMS USA: Version 10.0 [Dataset]." *Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0.* .

Growth of linked data (according to Web of Science...)



Correct Reference Population for Weighting

- What is the target population?
- Overlap between dataset A and dataset B
- Think about mortality selection and in and out migration