Population Research with Linked Data: Guide to Inference

SSHA 2024

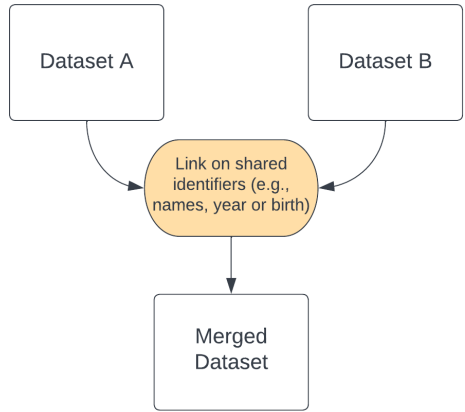
Casey F. Breen[1] Won-tak Joo[2]

December 7, 2024

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

Record linkage

- Identify same person across datasets in absence of a unique identifier (e.g., SSN)
- Wide applications: demography, sociology, computer science, epidemiology, history, medicine, economics, industry, etc.



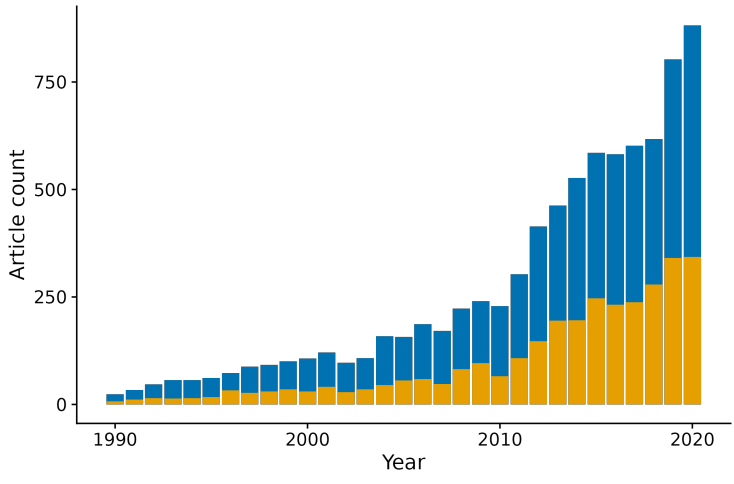
The growth of linked data in the social sciences

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

The growth of linked data in the social sciences

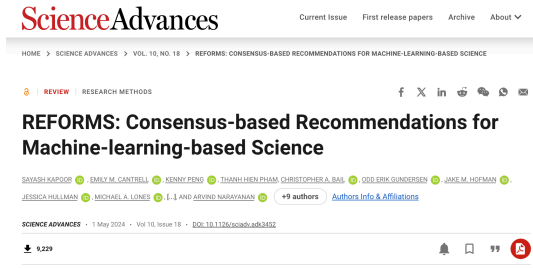
- Explosion in publicly-available linked census and admin data (Ruggles et al., 2020; Genadek and Alexander, 2022; Goldstein et al., 2021; Abramitzky et al., 2020)
  - Much lower barriers to entry
- Large and important body of methodological research on improving record linkage (Ruggles, Fitch and Roberts,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

Growth of linked data



Less methodological attention to inference

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



Example from machine learning...

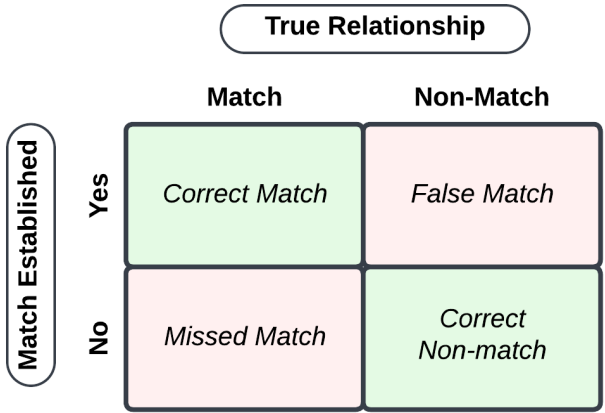
This study...

1. Framework for unpacking bias in estimates due to linkage errors
2. Checklist for inference with linked science

Framework for inference with linked data

- Two types of linkage **error** with distinct consequences for inference
  - *Missed Matches (Type II Error)*: Failing to link true matches.
  - *False Matches (Type I Error)*: Incorrectly linking different records.

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.

Non-Probability Toolkit

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

Correct reference population

- What's the target population?
- Overlap in dataset A and dataset B
- E.g., if linking 1900 and 1940 census must account for differential mortality

False matches — descriptive rates

- **No false matches**:

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

  - O = Count of events/outcomes
  - N = Total population size

False matches — descriptive rates

- **No false matches**:

$$R = \frac{O}{N} \tag{3}$$

  - O = Count of events/outcomes
  - N = Total population size

False matches — descriptive rates

$$R' = \underbrace{R_{\text{true}} \times (1 - f_r)}_{\text{Contribution of True Matches}} + \underbrace{R_{\text{false}} \times f_r}_{\text{Contribution of False Matches}} \tag{4}$$

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

False matches — regression coefficients

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{5}$$

where:

$$\hat{\beta}_1 = \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(X)} \tag{6}$$

False matches — regression coefficients

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

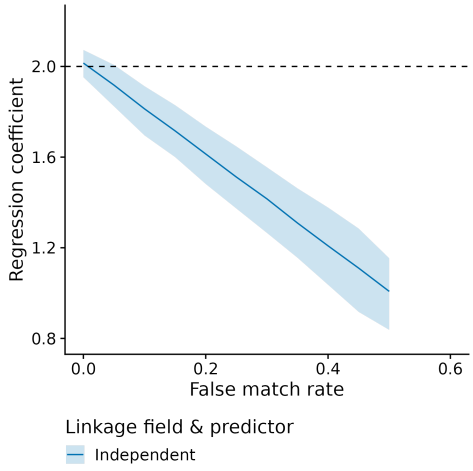
Regression framework: assuming no covariance in false matches

$$\hat{\beta}_1' = \frac{(1 - f_r) \cdot \mathsf{Cov}(X, Y) + f_r \cdot \cancel{\mathsf{Cov}(X_{\mathsf{false}}, Y_{\mathsf{false}})}}{\mathsf{Var}(X)} \tag{8}$$

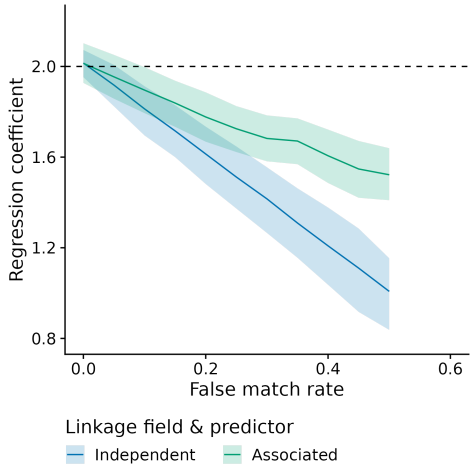
$$= \frac{(1 - f_r) \cdot \mathsf{Cov}(X, Y)}{\mathsf{Var}(X)} \tag{9}$$

$$= \hat{\beta}_1(1 - f_r) \tag{10}$$

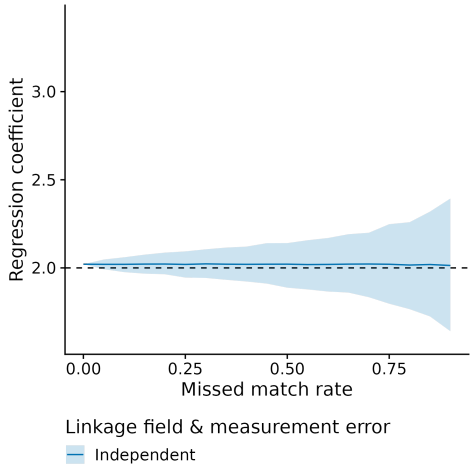
Simulation Results



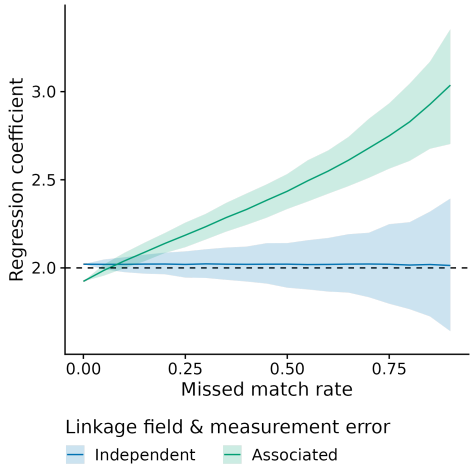
Simulation Results



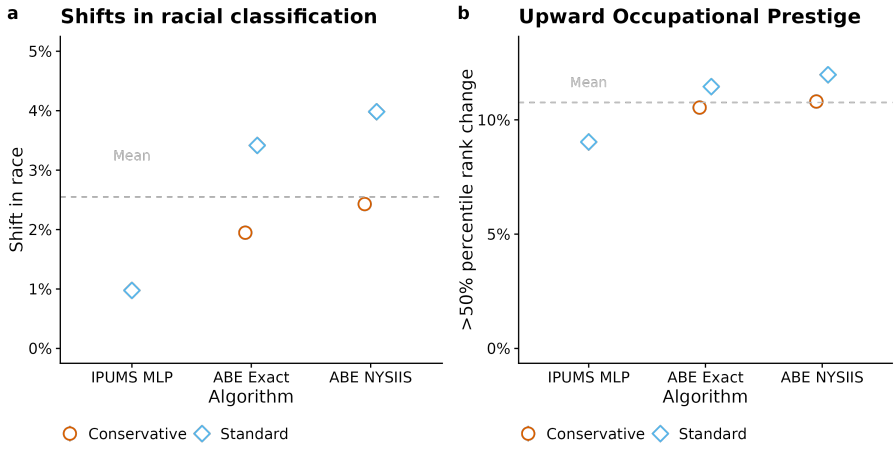
Simulation Results



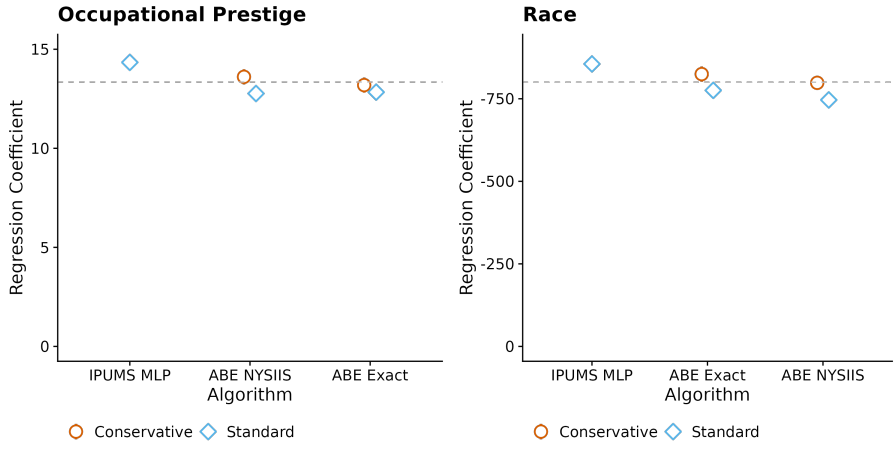
Simulation Results



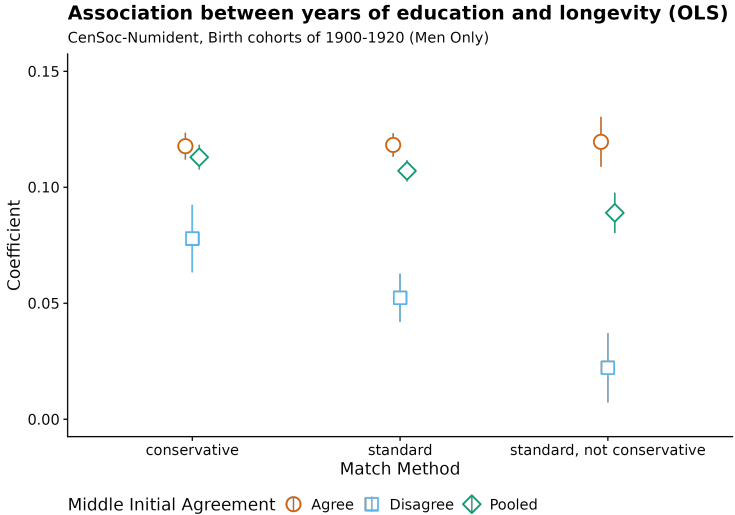
Empirical Results



Empirical Results — regression on wage/salary income



Empirical results — validation variable (middle initial)



Checklist for linked data

- Checklist for researchers, reviewers, and editors
- Help promote transparency and replicability in record linkage science

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

Checklist: Describe Linkage Approach

1. Describe linkage methods
   - *Clearly describe and justify linkage methods/algorithm used (e.g., deterministic, probabilistic), including linkage fields*
2. Report basic descriptives
   - *Report match rate, number of matches established, and any other relevant metrics.*
3. Ensure replicability
   - *Release code and data to replicate linkage (to extent possible)*

Checklist: Assess linked sample

4. Quantify Representativeness of Linked Sample
   - *Evaluate how representative linked sample is of the target population. Check whether findings are robust across different algorithms (if possible)*

5. Validate Linked Data
   - *Investigate whether a validation variable exists (e.g., middle initial) or another approach for quantifying match accuracy*

Checklist: Implications of Linked Sample

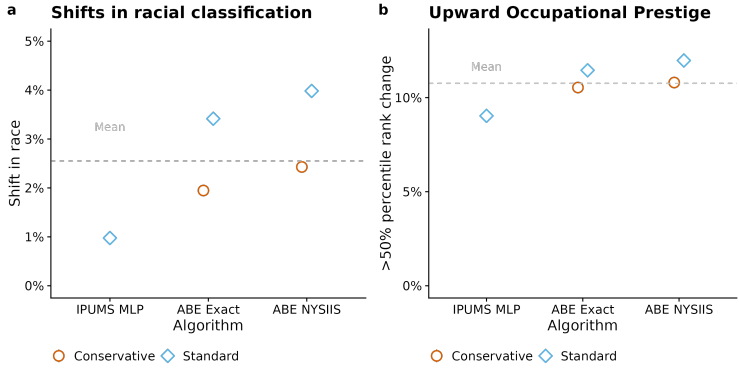
6. Report Implications for Research Results
   - *Discuss how linkage errors impact findings (coefficients attenuated? Rates upwardly biased?)*

7. Address Privacy and Ethical Concerns
   - *Ensure privacy measures are in place and ethical approvals are documented. Address all privacy and data protection concerns.*

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.* .