Adding Production Quality Race and Ethnicity to the LEHD Master Files

John Abowd, Kevin McKinney, and Lars Vilhuber

Overview

- Methodology
- Data Sources
- Work Flow
- Outputs
- Impact on QWI / OTM
- Status
- Conclusion

Methodology

- Direct match when available
- Statistical link otherwise
- Unified Bayesian approach for imputing all person specific characteristics (sex, DOB, race, and ethnicity).
- Create an estimation / training dataset using multiple data sources

Data Sources

- Person Characteristics File (SSA)
  - Primary source for sex and DOB. Limited race and ethnicity information (Black, White, Asian, no race for Hispanics) is also available.
- 2000 Decennial Census (Short Form)
  - Primary source for race and ethnicity
- American Community Survey
  - Secondary source for race and ethnicity
- Unemployment Insurance and ES-202
  - Work history, firm, and co-worker characteristics

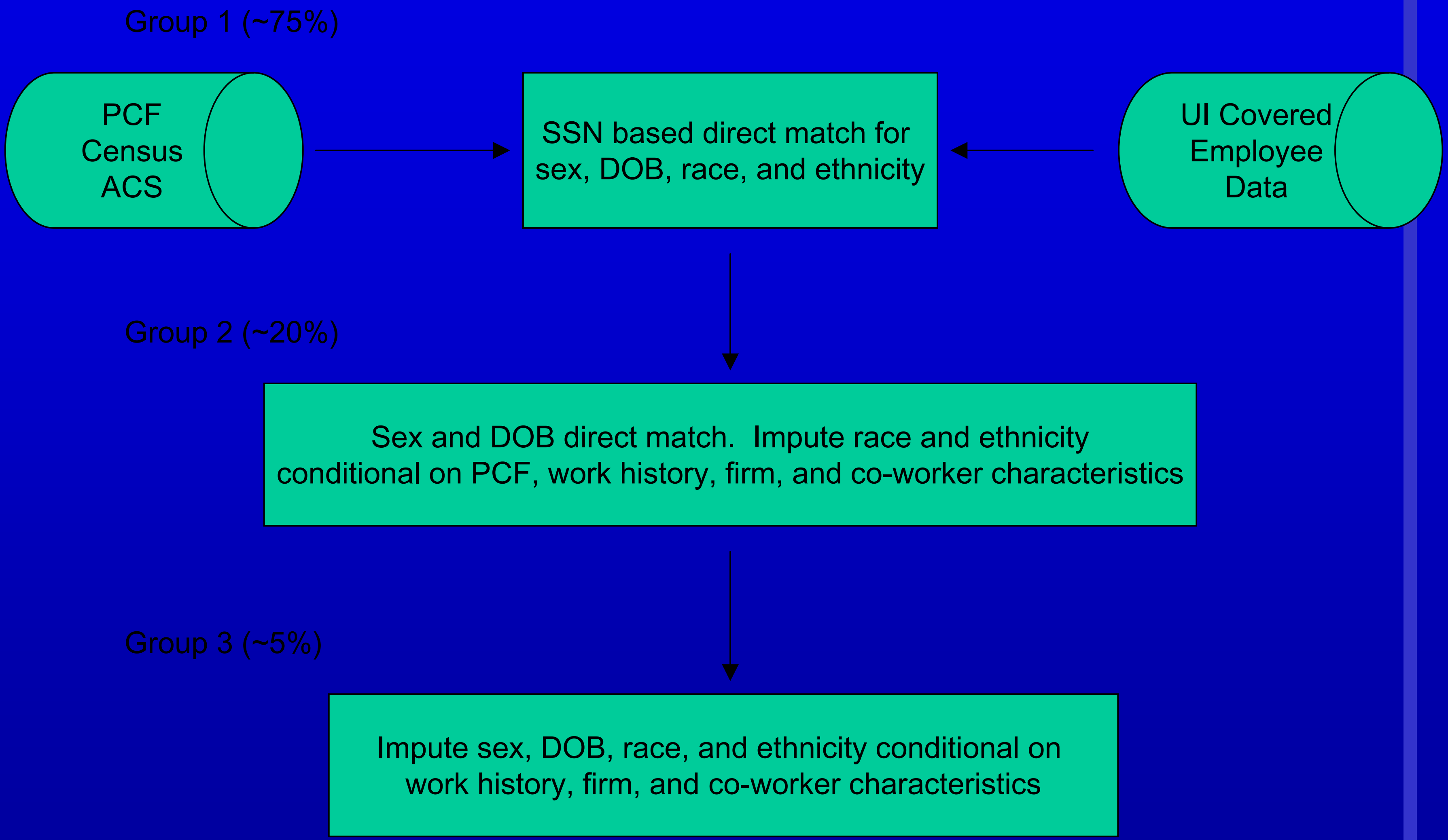
Methodology 2

- PCF and UI information are used to create cells of workers with the same observable characteristics
- Within each cell or cluster, a Bayesian kernel density approach is used to estimate the joint Posterior Predictive Distribution (PPD)
- An impute for a particular worker is a draw from the appropriate PPD.

Data Availability Drives Our Approach

- Group 1 (Direct Match 75%)
  - Sex and DOB from PCF. Race and ethnicity from ACS / Census.
- Group 2 (Partial Impute 20%)
  - Sex and DOB from PCF. ACS / Census Race and ethnicity imputed conditional on PCF race and ethnicity.
- Group 3 (Full Impute 5%)
  - No Census data sources available.

Work Flow



Group 1 (~75%)

PCF
Census
ACS

SSN based direct match for sex, DOB, race, and ethnicity

UI Covered
Employee
Data

Group 2 (~20%)

Sex and DOB direct match. Impute race and ethnicity conditional on PCF, work history, firm, and co-worker characteristics

Group 3 (~5%)

Impute sex, DOB, race, and ethnicity conditional on work history, firm, and co-worker characteristics

Outputs

- Sex and DOB
- Race
  - White
  - Black or African-American
  - American Indian or Alaskan Native
  - Asian
  - Native Hawaiian or Pacific Islander
  - Other Race
  - Two or More Races
- Ethnicity
  - Hispanic

Outputs 2

- Two or More Races
  - American Indian or Alaskan Native and White
  - Asian and White
  - Black or African-American and White
  - American Indian or Alaskan Native and Black or African-American
  - Multiple race combinations >1% of population
  - All other multiple race combinations

Impact on QWI / OTM

- Large expansion of current tables
  - Sex*Age = 2*8 = 16 cells for each county (metro, WIA), industry, ownership class
  - Sex*Age*Race*Ethnicity = 2*8*7*2 = 224 cells for each county (metro, WIA), industry, ownership class
- Potentially a large number of suppressions, especially for small size multiple race combinations

Status

- Multiple data sources have been integrated into an estimation data set
- Currently developing Bayesian methods compatible with our non-parametric approach
- Working towards first full-scale run in research environment.
- QA before handoff to production team

Conclusion

- Three fourths of records have race and ethnicity attached using a direct SSN link
- Only 5% of records have no person specific information on race and ethnicity
- Non-parametric Bayesian kernel density imputation used for workers without a direct SSN link
- First QWI products expected fourth quarter of 2010