Labor Composition and Quality Using Augmented CPS Data on Industry and Occupation

Peter B. Meyer and Kendra Asher

Office of Productivity and Technology, U.S. Bureau of Labor Statistics

ASSA conference, SGE session on Productivity Puzzles

January 4, 2021

Views presented by the authors do not represent views of their agencies

Context

- The Current Population Survey (CPS) gets monthly data from ~60,000 households
- Each job is assigned a Census-defined industry category and an occupation
  - These are 3-digit codes, used in the CPS, ACS, and other data sets
  - Challenge: The categories have changed over time

- We need long time series for industries and occupations
  - Our intended application: labor composition indexes by industry

- Past approaches: Crosswalks; or, study each category for customized imputation
- Approach here: Impute for each individual by machine learning

Census industries and occupations

- Hundreds of discrete groups, with 3-digit numbers

CPS period	Occupation categories	Industry categories
1982-92	394	229
1993-1999	456	237
2000-2010	503	264
2011-2012	533	263
2013-18	484	260

- Industry and occupation are coded (assigned) jointly
- Same classification is used in Population Census, CPS, and ACS
- Challenge: standardize comparison of observations across time & datasets
  - To follow one category over time
    - E.g. electrical engineers category grew and split, creating software categories
    - In our case, to fill in NAICS industry code consistently over time
  - To hold industry or occupation constant in a study of something else

Harmonizing industry and occupation over time

- A **crosswalk** or concordance matches the categories over time
  - It's a **table** where each category is mapped into categories in the other classification system
  - To avoid empty cells, destination categories may be merged
  - Users trade off precision of category with sparseness and length of time series
  - E.g. in 1960 there was one Census category for lawyers and judges

Counts of lawyers and judges in decennial Census 1% samples				
	1960	1970	1980	1990
Lawyers	2053	2570	5082	7603
Judges		123	298	331

- Researchers choose among crosswalks; there is a quiet literature on this
  - IPUMS (1994 and on), Meyer and Osborne (2005), IPUMS (~2007), Dorn (2009)

Examples from occupation study

- Lawyers and judges
  - Combined in 1960 Census of Population; separate in 1970-1990
  - Can split them apart?
  - Yes, using other Censuses as training data (1970-1990) and logistic regression
  - Predictors: State employee, Federal employee, Age, age^2, earnings cubic, business income
  - And: Education level below 16 years (!)

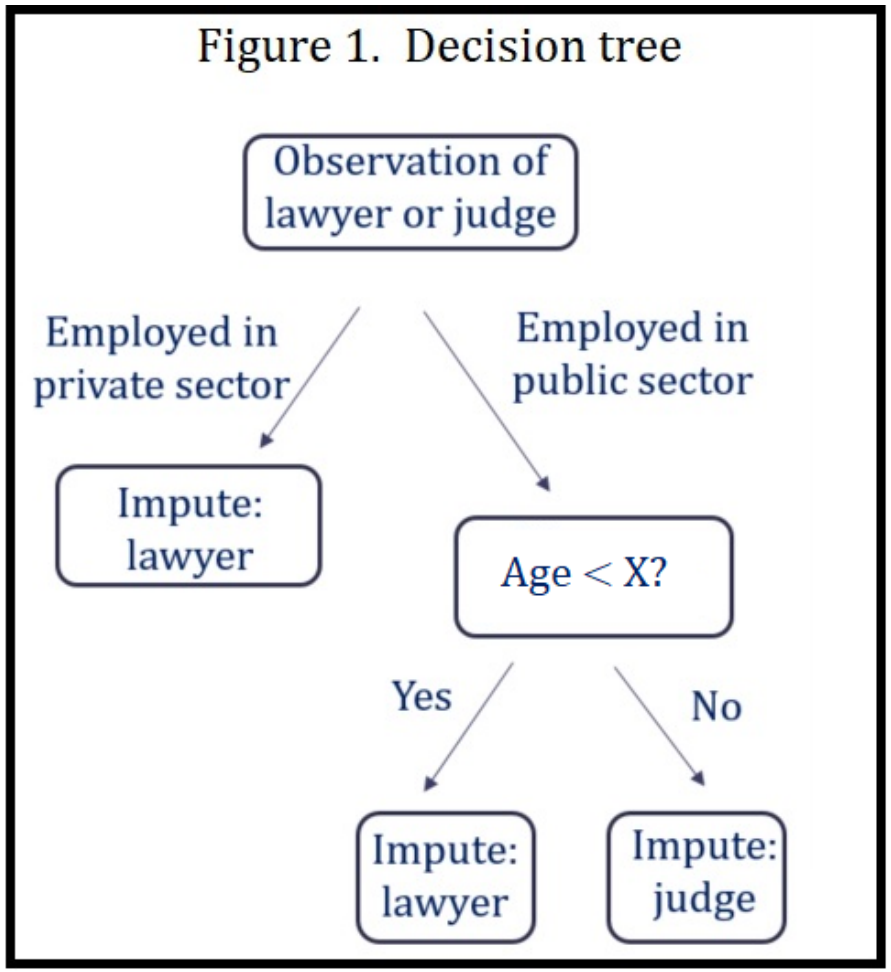
- Statisticians and actuaries
  - Again, combined in 1960 Census of Population; separate in 1970-1990
  - Predictors: Industry, Age, age^2, earnings cubic, education
  - And: Lives in CT, MN, NE, or WI

➔ Surprising categories and thresholds help predict, but it's too labor intensive

Random forests

The random forests method can make large scale imputations to individual records.

- We use the `ranger` package, which works well with many data types, notably categories
- Builds decision trees of linear combinations of predictors, threshold values, and category divisions – based on training data set
- Many decision trees "vote" to make a prediction in the test data set
- Random forests can incorporate interaction effects, and tend not to overfit.



Tuning the random forest classifier

Main tuning parameters for each imputed variable:

- Number of variables at branches of decision trees
- Numbers of trees

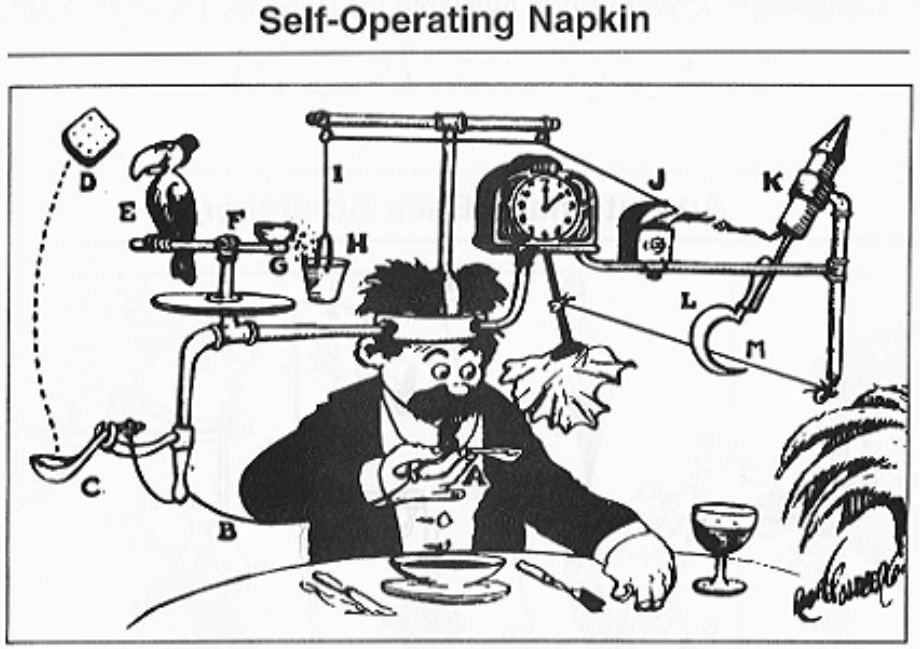
These are limited by computer memory and time.

Goal: High accuracy of out-of-sample predictions in the dual-coded data

Random forests models are complicated

- The "forest" is complicated to describe
  - Like a Rube Goldberg machine
- Variables' importance in prediction is available
  - Less info than from a regression

- We think random forests fit the problem well.
  - Logistic regressions are easier to check and control, but too labor-intensive
  - Neural networks may not give much additional benefit with this class of problem, and can require more resources.



Rube Goldberg, *Collier's*, Sept 26, 1931
From Wikimedia Commons

CPS and ACS data

Main training data set:

- Dual-coded CPS sample from 2000-2002
  - Dual-coded means it has **both** Census 1990 and Census 2000 industries and occupations
  - Coded by the Census specialists

Main target data: Monthly CPS 1986-99 combined with IPUMS-CPS

- 15.5 million observations

We also make imputations to:

- 2000-2018 CPS
- American Community Survey (ACS), 2003-2018

Several imputations are necessary

- Main goal: impute after-2000 industry to earlier microdata
- We train predictions on the dual-coded 2000-2002 CPS for:
  - Class of worker (e.g. for profit, not for profit, government)
  - Hours of work, attributes of any 2nd job
  - Occupation (2- and 3 digit Census 2000)
  - Industry (3 digit Census 2000)
  - NAICS industry

Work, location, and demographics predict industry

Strongest predictors:

- Industry (in earlier/native category system)
- Occupation
- State of residence

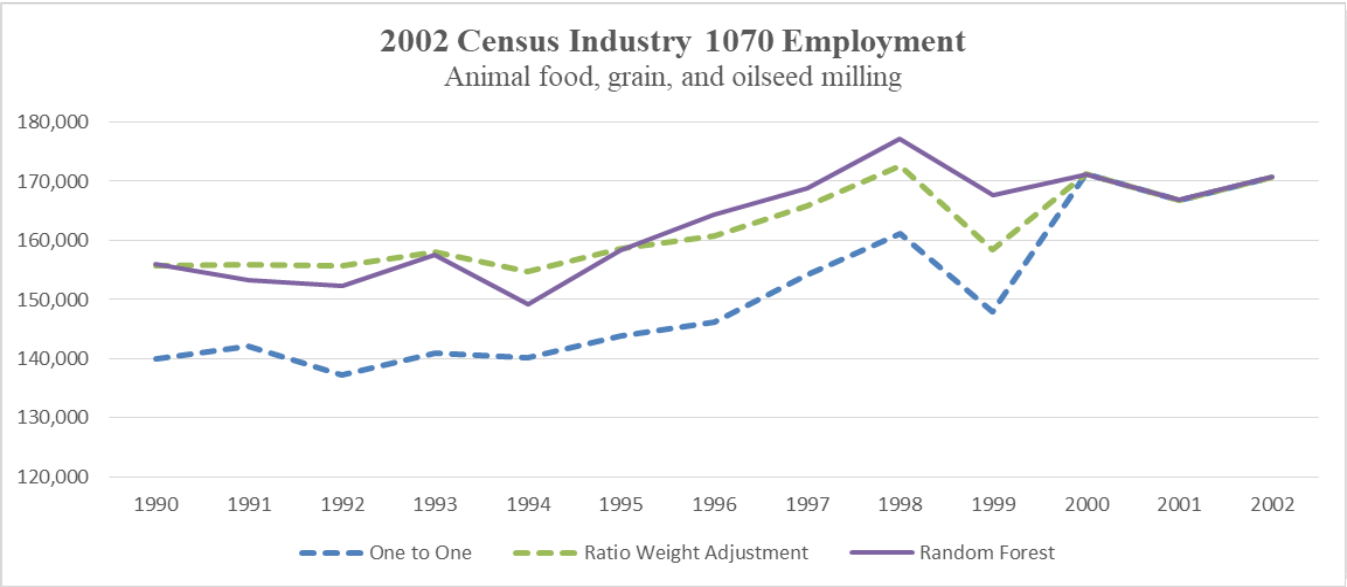
Weaker predictors:

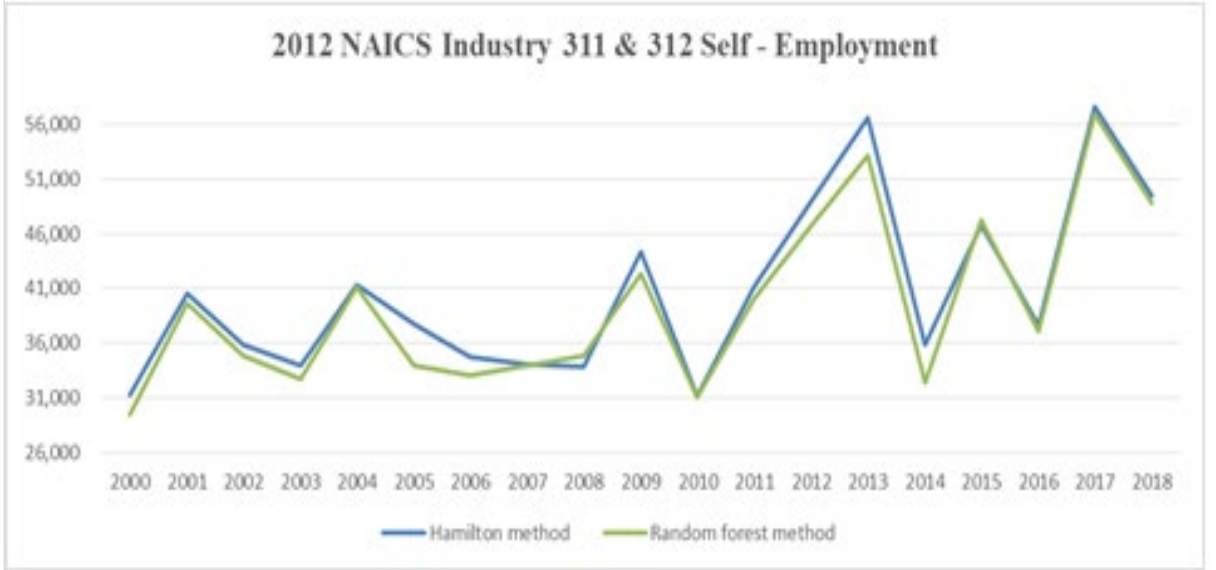
- Education, age
- Earnings, work hours, employer type
- Age, sex, race
- Metro, county
- State unemployment, from Local Area Unemployment Statistics

Challenge: Other variables definitions change in CPS notably in 1994 redesign

Creates an augmented CPS dataset

- We get imputations in an "augmented CPS" dataset for 1986-2018.
- To compare methods we estimate industry employment and self-employment
- Some imputations look good on the micro level.
  - Example: Durable vs nondurable manufacturing for "not specified manufacturing" industry 3990
- Milling industries were reclassified in 2000.
  - Imputations modestly change aggregates:





Benchmarks to apply

- Broad tests of the augmented data set are necessary
  - Imputations may be biased toward the conventional
- Benchmarks: Total in each industry and occupation
- Each occupation and industry category should evolve slowly
  - the fraction of the population in each category
  - average earnings in each category
  - demographic and geographic distribution

Application: Labor composition indexes

Our office has an established technique to create an index summarizing the education and experience of workforce in each industry (BLS, 1993)

- More educated and experienced workforce correlates to more output
- So the index accounts for some of productivity growth, apart from hours worked

- The index is constructed from data on individuals from the CPS
- For small-sample industries that gives a volatile index

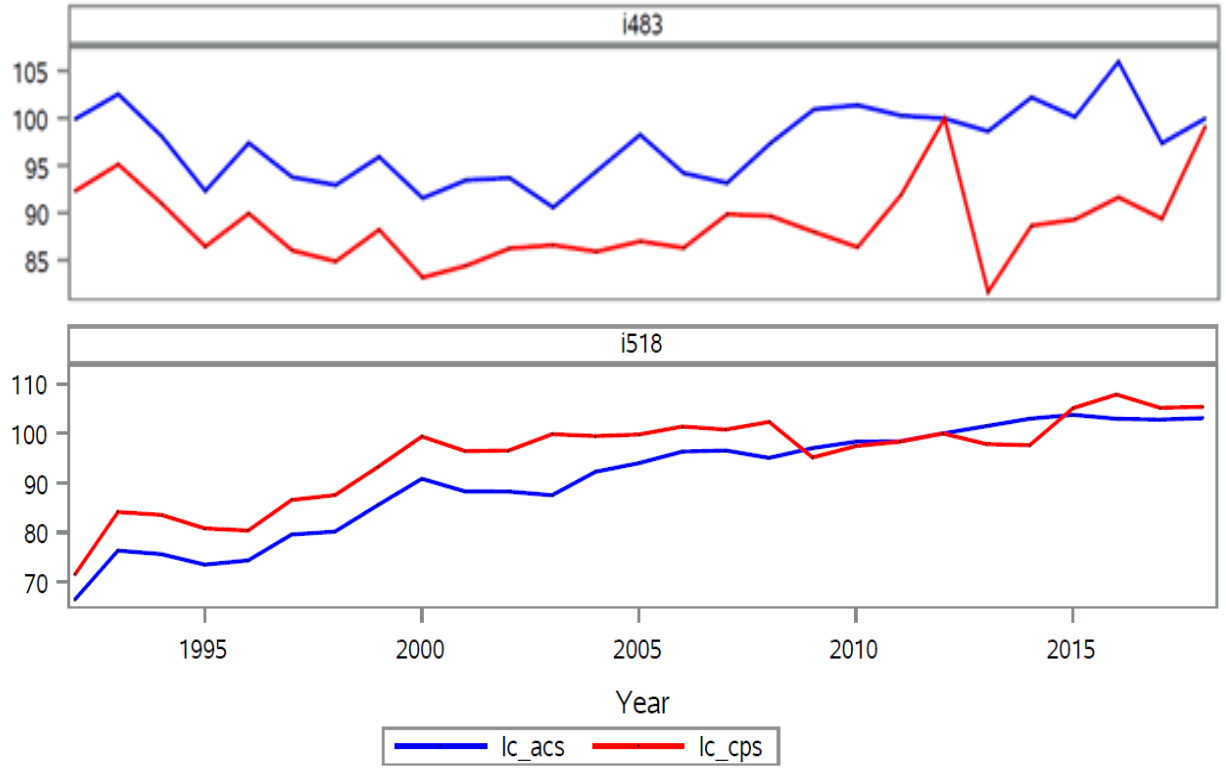
We'd like more accurate industry imputations

- Often smooths out fluctuations
- To create indexes for smaller industries

We test our imputation method on American Community Survey (ACS) data

Labor composition from ACS and CPS

- Indexes combine change in education and experience of industry workforces
- Indexes are 100 in base year 2012
- Red lines are CPS
- Blue lines are ACS-based
- Both raked to CPS three-year averages by 2-digit sector and age group, sex, and education groups
- Differences between them are from sample variation at the 3-digit level
- Larger sample size in ACS ➔ less volatility in these 3-digit indexes



NAICS 483 - Water Transportation

NAICS 518 - Data processing, hosting, and related services

Conclusions

The random forest approach gets us key benefits

- Large scale assignment of industry and occupation for CPS
- Using data on every person's primary job
- First known implementation of this
- Expected to be more accurate than a crosswalk, more feasible than logit regression

Long term

- use more data sets as training data (NLSY, population Census)
- apply to other data sets (older CPS and population Census)

Contact

Peter B. Meyer
Research economist
Office of Productivity and Technology
U.S. Bureau of Labor Statistics
Meyer.peter@bls.gov
bls.gov/dpr/authors/meyer.htm