Augmented CPS Data on Industry and Occupation

Peter B. Meyer and Kendra Asher
Office of Productivity and Technology, U.S. Bureau of Labor Statistics

GASP 2020 workshop

November 6, 2020

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

Context

- The CPS (Current Population Survey) 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

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

- Hundreds of discrete groups, with 3-digit numbers
- Industry and occupation are coded (assigned) jointly
- Same categories used in Population Census, CPS, ACS, and other data
- 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
  - Trade off precision of assignment with sparseness and length of time series

  - Industry example: "Animal food, grain, and oilseed milling" is new in 2000.
  - Occupation example: Lawyers and judges are sometimes categorized together
    - Can we separate them after the fact? Yes, pretty well, with micro data on each one.
    - Predictors: employed in public sector ; income ; age ; education thresholds

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

Scale up data and methods

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

Target data: Monthly CPS 1986-99 combined with IPUMS-CPS

- 15.5 million observations; we impute Census 2000 ind and occ

Random forests method for large scale of categories and data

- We use the `ranger` package, which works well with many data types
- Builds decision trees of threshold values and regressions in training data.

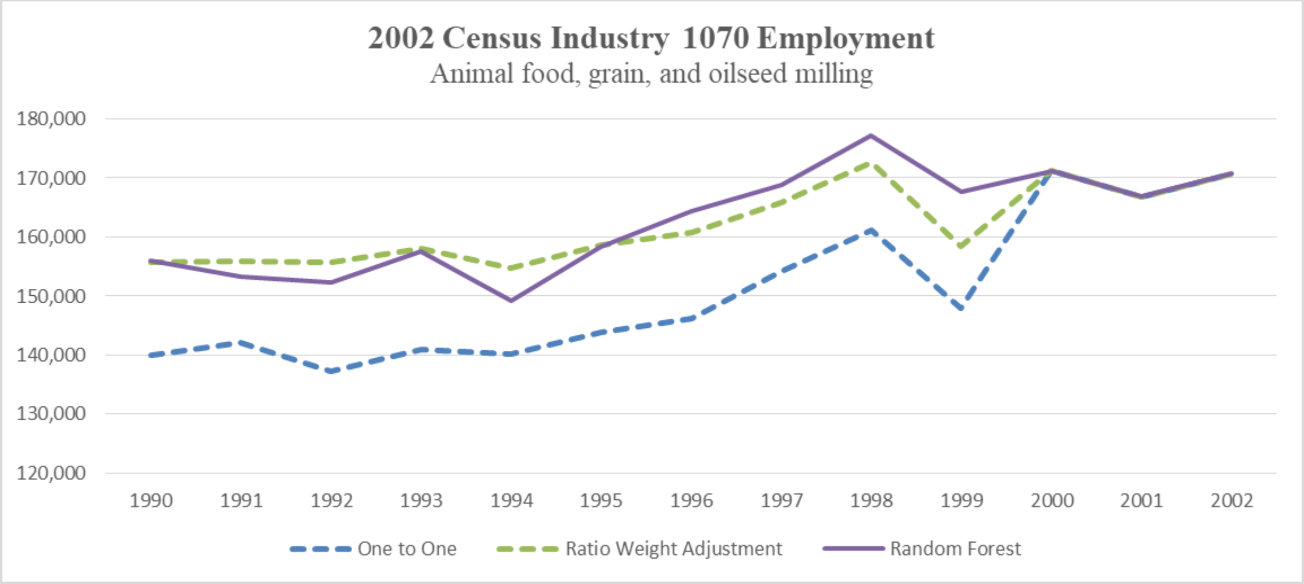
Several imputations are necessary

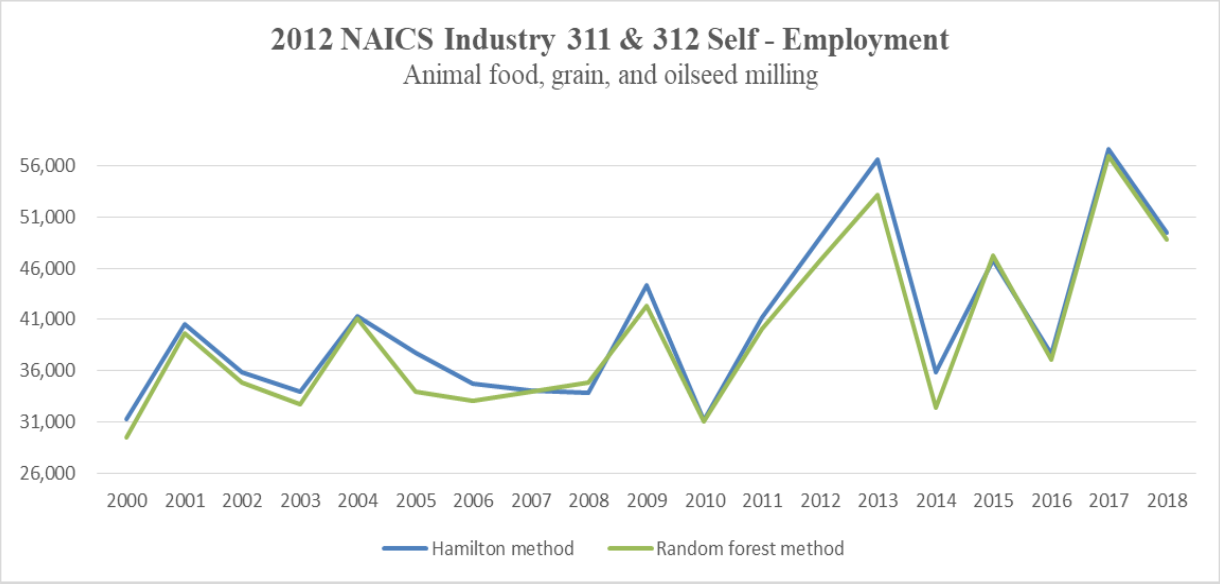
- We train predictions in the dual-coded 2000-02 data to impute:
  - Class of worker (e.g. for profit, not for profit, government)
  - Hours of work, attributes of any 2nd job
  - Occupation (3 digit Census 2000)
  - Industry (3 digit Census 2000), and NAICS industry

- Predictors of industry: work, location, and demographics
  - Strong: Industry (in earlier/native category system), occupation, state
  - Also education, earnings, work hours, employer type, age, sex, race, metro, county, year
- 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.
- We get **employment, self-employment,** & **work hours** estimates from this data
- Some imputations look good on the micro level. Examples:
  - Durable vs nondurable manufacturing for "not specified manufacturing" industry (Census 2012: 3990)
    - More data is usable after imputation.
  - This industry had classification changes in 2000, and our method modestly changes aggregates:





Benchmarks to apply

- Broad tests of the augmented data set are necessary
  - Imputations may be biased toward the “conventional”
- Benchmark: Total in each industry and occupation
  - Census 2000 totals (Scopp, 1993) – a macro test
- Each occupation and industry category should evolve slowly
- Can track time series of
  - the fraction of the population in category
  - average earnings
  - demographic and geographic distribution

Tuning the resulting classification

Tuning parameters for “classification forest” for each imputed variable:

- Number of variables at branches of decision trees
- Numbers of trees
- Proportional split between training and test sets
- Random seed

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

To match macro benchmarks:

- Can change thresholds in decision trees
- Multiple / fractional imputation, splitting respondents across imputed industries

Conclusions

The random forest approach gets us key benefits

- Large scale assignment of industry and occupation for CPS
- Using data on every person and job -- first known implementation
- Expected to be more accurate than a category crosswalk

More to do

- Test against benchmarks and adjust thresholds ; put to use in our applied problem
- More dual-coded data sets to use as input, and can impute to other data sets

Interested in advice and feedback

- Re industry and occupation coding, and
- On tuning parameters to random forest models

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