Augmented CPS Data on Industry and Occupation

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

Virtual JSM 2020
August 3, 2020

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

Outline

- CPS (Current Population Survey) uses industry and occupation codes that change over time
- We need consistent time series by NAICS industries for recent decades
  - Past approaches: Crosswalks; or, study each category
- New approach: Impute for each individual by machine learning
  - Training data: Dual-coded data sets
  - Random forests with `ranger`
- Tests and benchmarks to apply

Census industries and occupations

- Hundreds of discrete groups, with 3-digit numbers

- Industry and occupation are coded (assigned) at the same time
- Same categories used in Population Census, CPS, ACS, and other data

- Challenge: compare observations across time & datasets
  - To follow one category over time
    - E.g. electrical engineers category grew and split creating software categories
  - To hold industry or occupation constant in a study of something else
  - In our case, to fill in NAICS industry code consistently over time

Numbers of Census occupation categories	
1950	243
1960	296
1970	441
1980	504
1990	504
2000	543
2010	540
2018	569

Harmonizing industry and occupation over time

- A **crosswalk** or concordance matches the categories over time
  - It's a **table** where each category in one system is assigned to one or more categories in another
  - They can merge more or less, trading off precision and sparse-ness (empty cells)
  - No crosswalk will be best for all purposes
- Census Bureau regularly estimates how many people in previous categories would be in new categories, but does not impute this for each person.

- Key crosswalks, a partial history
  - IPUMS (1994-, from U of Minnesota Population Center) offers 1950 industry and occupation codes for any population Census or CPS observation
  - Meyer and Osborne (2005) applied 1990 occupations to 1960-2000 data
    - Shared that source code with ~50 people, but many empty occupation-year cells (sparseness)
  - IPUMS adopted that occ1990 and implemented ind1990
  - Dorn (2009) reduced number of MO's occupation categories to reduce empty cells

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; Zoghi, 2010)

  - 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

- For smoother indexes
- And to create indexes for smaller industries

Augmented CPS for this purpose means new column with NAICS industry implied by the data for each employed person.

Data sources

- CPS basic monthly files, with 15.5m observations
- IPUMS-CPS for 1986-1999
  - IPUMS imputes some variables we use
  - CPS redesign in 1994
- 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
  - One can use it for detailed study of a particular occupation's matches in other category systems
- Here we focus on imputing Census 2000 values to pre-2000 data

Example use of dual-coded data

We could study each occupation. Here, we predict 1990 occupation given 2000 occupation *within* the 2000-2002 dual coded data.

2000 category	1990 category	Predictors	In-sample accuracy
Farm, Ranch, Agricultural Managers	Farm managers	self-employed, older, high income	69%
	Farm workers	Private firm employee; age<21	
Appraisers and Assessors of Real Estate	Real estate sales	Self-employed ; Real estate industry	90%
	Public administrators	Public finance industry	
	Managers and administrators	Other industry	

But studying each occupation doesn't scale up, and would not meet economy-wide benchmarks.
Below we instead impute Census 2000 values to pre-2000 data on a large scale.

Several imputations are necessary

- Main goal: impute after-2000 industry to data from before 2000

- We train predictions in the dual-coded 2000-02 data to impute:
  - Class of worker (for profit, not for profit, government)
  - Hours of work, attributes of any 2nd job
  - Occupation (3 digit Census 2000/2010)
  - Industry (3 digit Census 2000/2010)
  - NAICS industry needed for our final productivity estimates

- Predictors of industry: work, location, and demographics
  - Most importantly: Industry (in earlier category system), occupation, state
  - Also: education, earnings, work hours, employer type/class, age, sex, race, metro, county, year

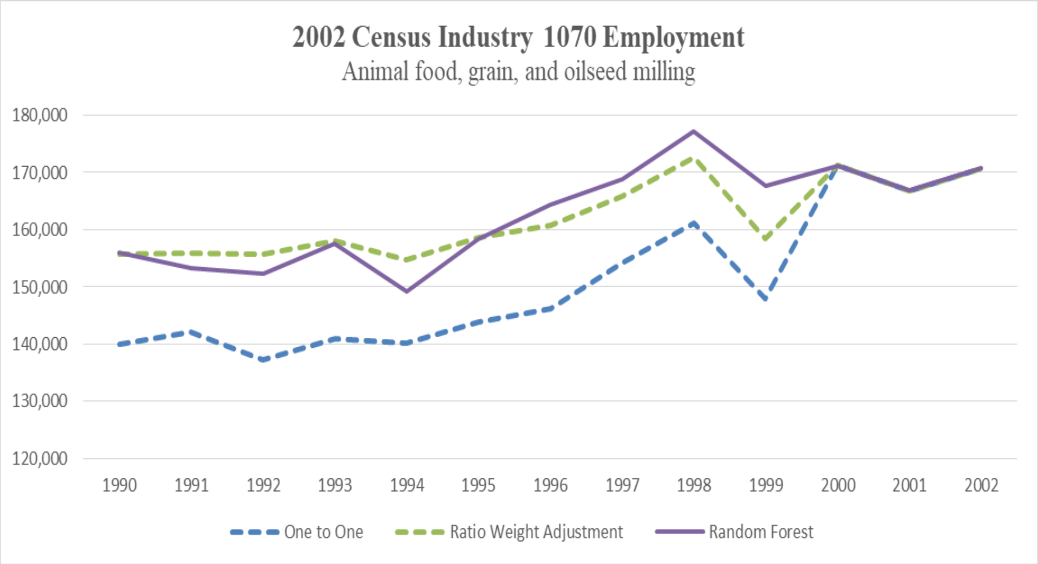
Random forests method

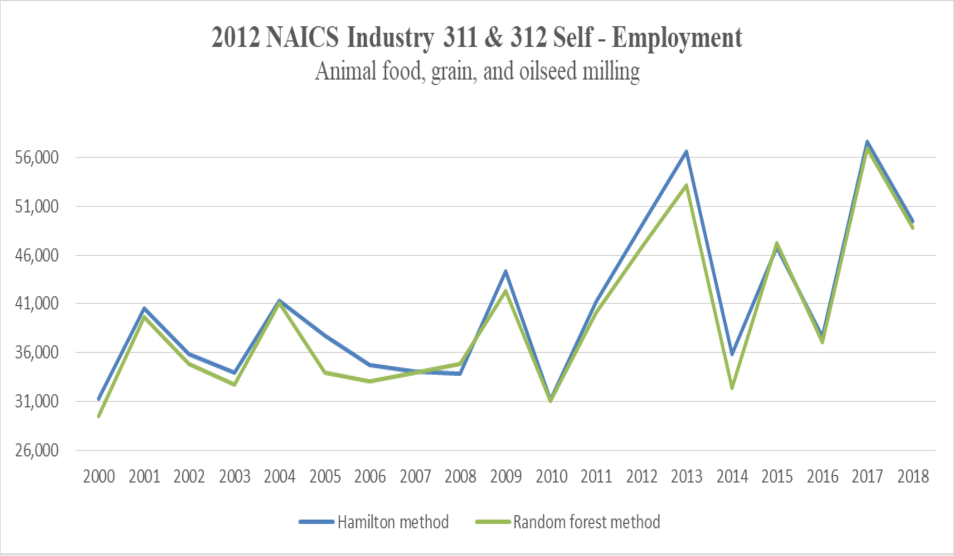
- Builds decision trees of threshold values and regressions in training data.
  - Automatically ; not studying each case
- There are several implementations of random forests in R

- We use the `ranger` package
  - Works well with many types of categorical data, other data types
  - Uses memory to the max, and time; hard to diagnose out-of-disk-space

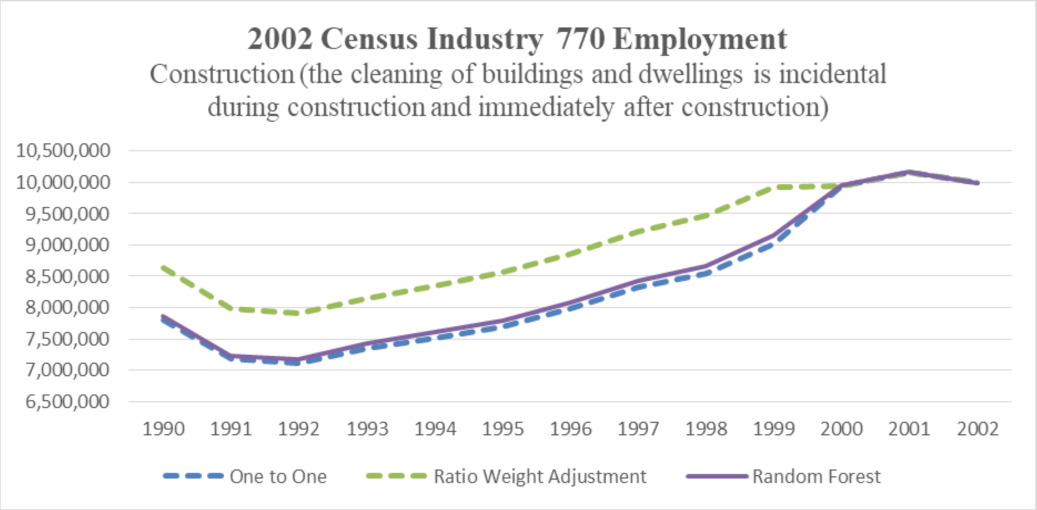
Creates an augmented CPS dataset

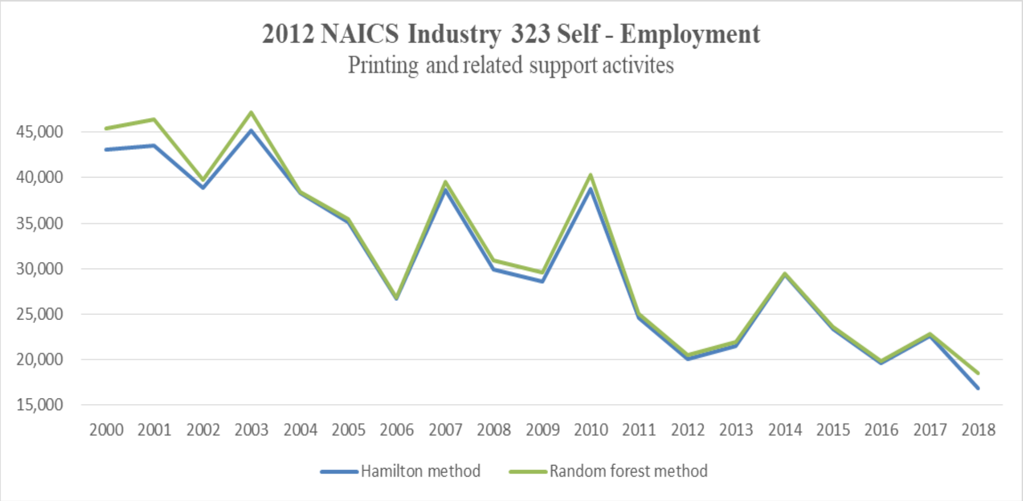
- We get imputations in an "augmented CPS" dataset for 1986-2018.
- Some imputations are good. Examples:
  - For each respondent in the category "not specified manufacturing industry" (Census 2012: 3990) we classify whether they are in durable manufacturing or nondurable manufacturing.
  - 1990 Census Ind 110 , "Animal food, grain, and oilseed milling" splits into later categories
  - We get employment, self-employment, & work hours estimates that include these workers





Estimates from augmented CPS dataset



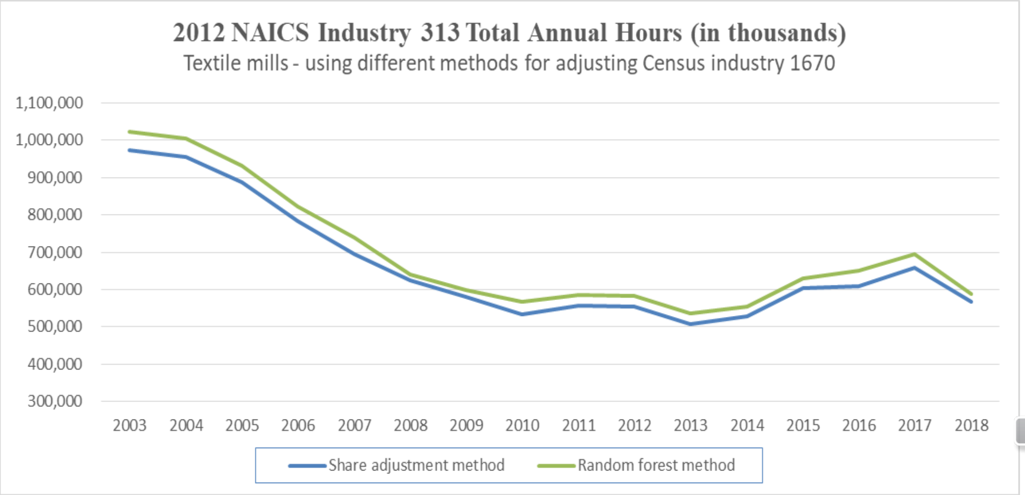


Can compute employment, self-employment, and hours worked from augmented CPS

We can compare them to the estimates from a crosswalk, or a "proportional reallocation."

In these cases the estimates are close to the old ones.

For details see Asher, Meyer, Varghese, (2019) and Meyer and Asher (2019)



Testing and benchmarking

- Broad tests of the augmented data set are necessary.

- Total in each industry and occupation in other sources
  - Census 2000 totals with analysis in Scopp (1993)
- Each occupation and industry category should evolve slowly
  - Can track time series of (a) the fraction of the population in category; (b) average earnings; (c) earnings variance; (d) demographic and geographic distribution.
- Imputations may be biased toward the "conventional"

Iterating to meet benchmark

Can adapt by

- changing thresholds on imputations
- add randomness to probabilistic assignments to “reinflate” variance

Multiple / fractional imputation may help

- Creating “fractional people” in synthetic population, splitting person-weights
- Impute both most likely industry and 2nd most likely industry, with probabilities

Extensions

More sources of external/dual-coded industry and occupation data

- 1970-1980 Census category change
- NLSY (National Longitudinal Survey of Youth) data
- Population Censuses can impute some things to the CPS

More data sets to augment:

- Augment Population Censuses and ACS with same methods

Conclusions

The random forest approach works and gets us key benefits

- Large scale assignment of industry and occupation for CPS
- Without analyzing each case
- Using individual information on each person, and
- Big data from other respondents and data sets
- It’s the first implementation I know of to do this

There’s more to do

- Test against benchmarks and adjust thresholds
- Create labor composition indexes with the new data

Expected to be more accurate than a category crosswalk

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