Improving Census to NAICS industry matches

Kendra Hathaway Asher, Peter B. Meyer, and Jerin Varghese
Office of Productivity and Technology, Bureau of Labor Statistics

- **BLS/OPT makes productivity estimates by aligning industry output to hours-worked in NAICS categories.**
- **We use Current Population Survey (CPS) microdata to supplement estimates of employment and hours worked, and enables us to adjust hours data for changes in workforce education and experience.**
- **CPS uses the Census industry classification systems, which we must map to 3- or 4- digit NAICS codes.**
- **Key issues: (1) assigning observations from "not specified" Census industries and**
- **(2) assigning observations from Census industry codes that map to multiple NAICS industries.**
- **Here we apply different algorithms for classification, and find that using geography, age, and occupation microdata improves match accuracy, and doesn't much change industry trends shown.**

For Data Linkage Day, Oct. 18, 2019 at NAS. Findings are preliminary. Views are those of the authors only.

(1) Not Specified Census industries

2012 NAICS	2012 Census Code	Census 2012 Category Title
Part of 21	480	Not specified type of mining
Part of 22	690	Not specified utilities
Part of 311	1290	Not specified food industries
Part of 331 and 332	2990	Not specified metal industries
Part of 333	3290	Not specified machinery manufacturing
Part of 31, 32, 33	3990	Not specified manufacturing industries
Part of 42	4590	Not specified wholesale trade

Number of employed observations in not specified Census industries

2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1,513	1,407	1,809	1,597	1,643	1,821	1,725	1,942	1,876	2,166	1,929

Hamilton's method for "Not specified" industries

Hamilton's method is used to apportion discrete numbers among a group. OPT uses the Hamilton method to allocate self-employed and unpaid family worker hours and employment from not specified industries to the component CPS industries. The method ensures component industries with larger employment shares receive more of the allocation, while controlling to the original aggregate industry's total employment and hours.

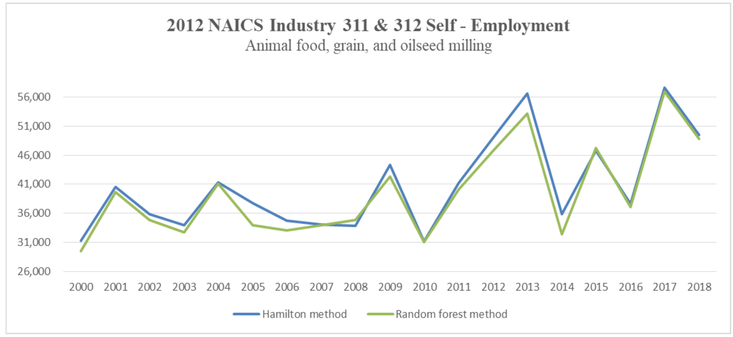
For labor composition estimates, CPS observations from Census industries 480, 2990, and 3990 are excluded. Their weights are indirectly added to other observations through benchmarking. The other not specified industries do not affect labor composition estimates.

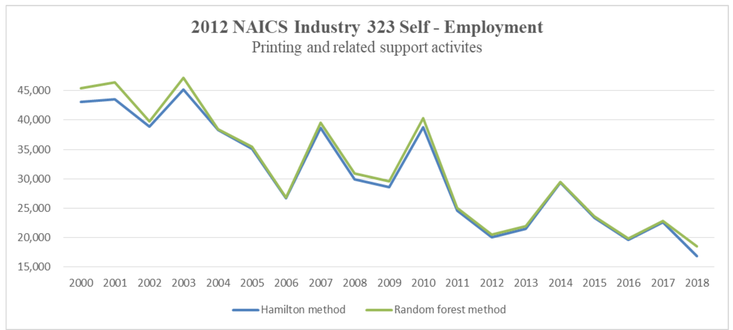
(2) Split industries by proportions

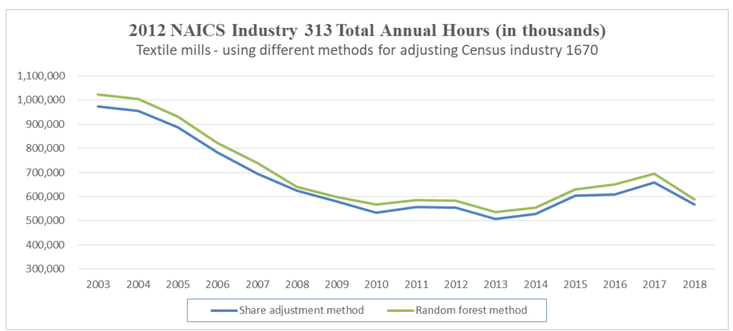
OPT uses various data sources to calculate industry shares used to split the employment and hours of Census industry codes into more detailed NAICS industries. These measures include number of establishments data from the County Business Patterns (CBP), number of partnerships from the Internal Revenue Service (IRS), all establishment sales from the Annual Retail Trade Survey (ARTS), sales from the Economic Census, number of paid employees from the Current Employment Statistics (CES), total receipts or total number of establishments from Non-Employer Statistics (NES), and revenue from nonemployer firms from the Service Annual Survey (SAS). Sources vary by industry depending on availability. Multiple data sources are used for some industries.

For labor composition, OPT divides each CPS observation into two and uses shares of hours worked to proportionally allocate the weights between the two industries. The total employment and hours of all observations are then summarized by industry and benchmarked to OPT's employment and hours measures by type of worker.

CPS annual self-employment and total hours using the two methods







(1) Random forest method for "Not specified" industries Example: 2012 Census industries 2990 and 3990

The first step for allocating not specified observations in Census Industry 3990 and 2990 is to determine whether observations in Census Industry 3990 belong to durable or nondurable industries. After homogenizing the variable definitions across year, the data are divided into groups based on Census industry, occupation, and year. Random forest algorithms are then trained on the other observations from manufacturing industries based on demographic, geographic, and occupation factors to predict each observation's manufacturing sector. Predictive accuracy reaches 94 - 96%.

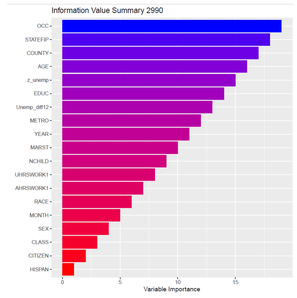
Next we create algorithms to predict the NAICS industry for both Census industries 2990 and 3990. Observations in Census Industry 3990 are subjected to their predicted Sector's NAICS industries. Observations in Census industry 2990 can only be predicted to be in nondurable goods NAICS industries.

2012 Census Industry 2990 (2014-2018)

```
Confusion Matrix and Statistics

          Reference
Prediction  331  332
       331 1519   60
       332  256 4189

    Accuracy : 0.9475
      95% CI : (0.9416, 0.953)
No Information Rate : 0.7053
P-Value [Acc > NIR] : < 2.2e-16
```



(2) Random forest method for split industries 2012 Census Industry 1670

Census Industry 1670 observations match both NAICS Industries 313 and 315. Algorithms can help assign observations between the two industries. We divide the Census 1670 observations based on each one's similarities to other observations in industries 313 and 315. We train our algorithm based on these other observations to predict which NAICS industry to assign to observations from Census industry 1670.

2012 Census Industry 1670 (2014-2018)

```
Confusion Matrix and Statistics

          Reference
Prediction  313  315
       313  954   54
       315   42 1023

    Accuracy : 0.9537
      95% CI : (0.9437, 0.9623)
No Information Rate : 0.5195
P-Value [Acc > NIR] : <2e-16
```