Augmented CPS Data on Industry and Occupation

Peter B. Meyer, Kendra Asher
Office of Productivity and Technology
U.S. Bureau of Labor Statistics
Nov 7, 2020[1]

Abstract
The Current Population Survey (CPS) classifies the jobs of respondents into hundreds of detailed industry and occupation categories. The classification systems change periodically, creating breaks in time series. Standard concordances bridge the periods, but often leave empty cells or inaccurate sharp changes in time series. They also usually build in the assumption that categories from a certain period of time can be representative, on more aggregate levels, and of longer historical periods.

For each employed CPS respondent from before the year 2000 we impute post-2000 Census industry and occupation classifications and related variables. The imputations use micro data about each individual and training data sets that were classified by specialists into two industry and occupation category systems – that is, they are dual-coded.

We train a random forests classifier to handle the changes in classification between the 1990s and 2000s largely on the dual-coded data set and apply it to the full CPS and IPUMS-CPS to impute several variables, including industry and occupation. For changes in classification when an industry or occupation splits, we train the algorithms on the observations with the newly classified industry or occupation split to predict how the historical observations would have been classified. We generate an augmented CPS, with additional columns of standardized industry and occupation. This data can serve research on many topics.

Keywords: CPS, prediction, imputation, occupation, industry, classification, employment

1. Introduction

The Current Population Survey (CPS) classifies the jobs of respondents into hundreds of detailed industry and occupation categories. The classification systems change periodically, historically each decade at the time of the Census of Population. This creates breaks in time series. Standard concordances bridge the periods, but often leave empty cells or artificially sharp changes in time series. For estimates about the composition of the workforce by industry, researchers want

[1] Views presented by the authors do not represent views of the Bureau of Labor Statistics. This paper was presented at the 2020 Joint Statistical Meetings in the Government Statistics Section's Session on Survey Weighting, Imputation, and Estimation.
https://ww2.amstat.org/meetings/jsm/2020/onlineprogram/AbstractDetails.cfm?abstractid=313878. New versions will be presented at the GASP (Government Advances in Statistical Programming) workshop on Nov 6, 2020, and at the ASSA conference Jan 3-5, 2021.

smoother time series for industry and occupation. In our production application, the BLS's multifactor productivity estimates depend on microdata from the CPS to supplement estimates of employment and hours worked, and to adjust hours-worked data for changes in worker education and experience.

For each employed CPS respondent from 1986 to 2018, we have applied prediction methodologies to impute standardized industry and occupation categories, employer class, and some estimates of hours worked, for each job. The imputations use micro data about each individual and large training data sets about the population. In some of the training data sets, industry and occupation have been classified by specialists into two industry and occupation category systems – that is, they are dual-coded. This project can help analyze the time series around the definitional classification breaks and smooth out their effects. Similar techniques will apply to the American Community Survey (ACS) and the decennial Censuses up to 2010, to help smooth time series within these data sets and to match results from each to the others.

One way to map industries and occupations between systems is with a crosswalk table that matches match whole categories from one system to whole categories in another. In this paper we improve on the accuracy of a crosswalk approach by using information on each individual observation. Our new method imputes a standardized industry and occupation to everyone in the CPS data with machine learning. We use a statistical classification method called random forests, trained on special data sets which have been classified in more than one way. That gives us a data set, which we call an "augmented CPS." In the long run we want to test the industries and occupations in the resulting augmented data sets for smooth population proportions and wage levels and for how well they match known trends, benchmarks, and other data sources.

2. Industry and occupation codes in the CPS and NAICS

The industry and occupation categories in the Population Census and therefore the CPS are recorded as three-digit codes. Each job is coded into one industry and one occupation. Electrical engineers, for example, were category 12 in the 1970 Census, 55 in 1980 and 1990, and then were split into 140 and 141 in the 2000 Census. The classifications have gained detail over time, with 296 occupations in the 1960 Census and 543 in the 2000 Census. The industry and occupation codes are filled in by specialist coders at the Census Bureau based on short descriptions of the person's employer and work tasks.[2]

The relevant CPS data sets are large and widely used. The main public use sample of the 1990 Census has observations of 6.5 million employed persons. The monthly CPS covers employed persons in about 60,000 households each month. Industry and occupation codes have three digits each.

[2] Meyer interviewed some of the "coders" who assign these industry and occupation codes in 2006. They were Census employees in a large facility in Jeffersonville, Indiana. The information they have was usually a text string describing job title, tasks, or responsibilities. They often had some version of the employer's name from the respondent. They could try to look up the employer in a variety of reference works or on the Web. They would generally fill in a code for industry first. That would help them pick occupation. The computer system would offer likely choices based on the text or other codes.

Occupation and industry classifications in these data sets are used in a variety of ways in social-scientific research. First, researchers construct estimates for the population in each category. Second, researchers often hold occupation and industry categories constant with a fixed-effects estimation in order to study something else. For example, studies of inequality study how much is happening within these categories or between them, as is done in Autor and Dorn's (2013) labor polarization study. Third, people use these categories to make estimates in other data and to impute or match them to data with Census information.

Data sets with these categories are large and widely used in social science research, but it's a challenge to make long time series and match them to other data. Working around classification breaks is a general problem, which statistical agencies confront regularly. If better methods for addressing classification changes can be developed, the accuracy of results from many studies would be improved.

The classification systems used in the CPS for industry and related variables change over time. Our office needs a consistent classification system to construct a time series of labor composition indexes by industry. We have used a couple of techniques to bridge changes in classification systems: simple crosswalks and, for the classification break between the 1990s and 2000s, proportional assignment to match aggregates. In this project we examine a new technique that we expect to be more accurate than the current methods. We impute consistently-defined industry, occupation, hours of work, and related variables to each individual CPS observation from 1986 to 2018. We make these imputations based mainly on a special "dual-coded" CPS sample which was coded into multiple classification systems around the time of the 2000 Census. Our goal is to impute the Census categories used in data after 2000 to the prior data.

Table 1: Numbers of occupation and industry categories each year in our CPS data

CPS year	Occupation categories	Industry categories
1986	389	228
1987	388	226
1988	391	227
1989	389	228
1990	393	228
1991	394	229
1992	392	236
1993	456	236
1994	456	236
1995	453	236
1996	451	236
1997	455	237
1998	449	236
1999	455	237
2000	503	259
2001	502	259

Table 1 shows the different occupation and industries in our CPS dataset by year. As seen in the table, over time the categories change in meaning, and the number of categories has increased. Major revisions, usually from the population Census, were introduced in the occupation categories in 1993, 2003, and 2013. The industry classifications changed substantially in 1993, and 2003, and modestly in 2012, 2014, and 2020. Both occupation and industry were recoded – dual-coded – for 2000-2002. Some fluctuations in the numbers between years occur in our data due to no observations with that classification being sampled in a particular year.

The CPS was substantially redesigned in 1994, though not in ways that affected industry and occupation classification (Polivka and Miller, 1995). Observations before then give us less data on certain variables, notably because they do not have data on second jobs.

2002	503	259
2003	503	264
2004	502	264
2005	501	264
2006	501	264
2007	502	264
2008	503	264
2009	503	263
2010	502	263
2011	533	263
2012	532	263
2013	484	260
2014	484	260
2015	484	260
2016	484	260
2017	484	260
2018	484	260

3. Crosswalks and their limitations

A standard way to create a longitudinally consistent classification system is the crosswalk. A crosswalk or concordance matches the categories over time. A crosswalk often uses a one-to-one table, and converts each industry or occupation under a previous classification system into one category in the new classification system. This is often considered a cruder method for conversion as often the two classification definitions are not the same. If two people have the same industry in one classification system, they will be classified the same way after a crosswalk under the second system, often regardless of other differences.

A crosswalk can be shown as a table in which each industry in a 1990s list is matched to one or more industries in the 2000s system. Analogously one could show a table of industries in which each industry in the 1990 Census list is assigned to one or more industries in the 2000s system. The designer of a crosswalk may need to groups some destination categories to trade off some precision to reduce sparseness. That is, a crosswalk that includes every detailed occupation will also have empty cells, which makes some comparisons or econometrics more difficult. No single crosswalk is best for all purposes.

For example, see the empty cells in these rows from a crosswalk of occupations (from Meyer and Osborne, 2005), which was designed to group some categories to make long term comparisons from the 1960s through 2010 possible:[3]

Table 2: Numbers of occupation and industry categories each year in our CPS

[3] These rows are from Meyer and Osborne (2005). Scopp (2003) has definitive lists of mappings between 1990s and 2000 Census and industry occupations.

Proposed standard job title	**Proposed standard code**	Census 1960 codes	Census 1970 codes	Census 1980 codes	Census 1990 codes	Census 2000 codes
Computer systems analysts and computer scientists	**64**		4; 5	64	64	100; 104; 106; 110; 111
Operations and systems researchers and analysts	**65**		55	65	65	70; 122
Actuaries	**66**		34	66	66	120
Adjusters and calibrators	**693**			693	693	
Water and sewage treatment plant operators	**694**			694	694	862
Power plant operators	**695**	701	525	695	695	860
Plant and system operators, stationary engineers	**696**	520	545	696	696	861
Other plant and system operators	**699**			699	699	863
Lathe, milling, and turning machine operatives	**703**	452	454; 652; 653	703; 704; 705	703; 704; 705	801; 802
Punching and stamping press operatives	**706**		656	706	706	795
Rollers, roll hands, and finishers of metal	**707**	513	533	707	707	794
Drilling and boring machine operators	**708**		650	708	708	796

Classification of workers into industries and occupations can be ambiguous, and also involves error. Scopp (2003) is the definitive work on the changes from the 1990 Census industry and occupation classification to the 2000 one. He wrote (page 9) that "any coding process involves coding error. In both censuses, these errors average about 7-8 percent for detailed industry codes, and 10-12 percent for occupation codes. These errors contaminate the comparisons across classifications, because they create false combinations of 1990 and 2000 codes." We infer from this that it is unrealistic overall to get better than about 7% error in industry assignment and 10% in occupation assignment, even when specialized coders are doing the work.

The Census Bureau regularly estimates how many people in previous categories would be in new categories, but does not impute this for each person. A number of efforts have been made to assign consistent assignments over time. Here are some we are aware of.

- 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 a simplified set of 1990 occupations to 1960-2000 data, using occupation titles mainly to match
- IPUMS adopted that definition for occ1990 and implemented ind1990 independently
- Dorn (2009) reduced number of Meyer and Osborne's occupation categories to reduce empty cells, and used a version of this system in Autor and Dorn (2013).
- IPUMS offers occ2010, an assignment of 2010 occupations to historic Census data.

4. Dual-coded data sets as training data

"Dual-coded" data sets are monthly CPS samples in which the industry and occupation have been coded into two different Census category systems. The classification has been done by the same specialists who would normally classify such observations. At the time the 1980 Census was conducted, a sample of 122,000 observations was dual-coded into both 1970 and 1980 classifications, and CPS samples from 2000 to 2002 were dual-coded. We have these data sets. (Meyer, 2010; Meyer and Asher 2019)

The key dual-coded data set used here was created to cover the change in industry and occupation between the 1990 and 2000 Census classification systems. This data set dual codes CPS monthly observations in 2000-2002. It has 2.4 million observations, with overlap as households were surveyed repeatedly. In this paper we call this the "bridge" data set because our main inferences are from the 1999-2000 changes.

A dual-coded data set can be used to study any particular category, or every category in turn. For example it is possible to predict (impute) 1990 Census occupation given 2000 occupation within the 2000-2002 dual coded data. In the next tables are examples from such a study. (Meyer, 2010) Accuracy can be estimated within the dual-coded data set itself, then the resulting coefficients or other model can be applied to CPS data in other years. This improves on what a crosswalk can do by taking more information into account.

Table 3: Imputation accuracy using logistic regression to impute 1990 occupation

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	

In principle it would be possible to conduct such studies of each industry and each occupation to use the dual-coded data to fill in smart imputations for every individual. In practice this is too much work. Meyer (2010) showed this was feasible for about 10 occupations, but a careful study of each occupation took days, and to cover the entire economy would take many person-years. Without covering most of the workforce, the imputation procedure can't benefit from known economy-wide benchmarks. That is, it cannot be calibrated to economy-wide benchmarks, and it will miss them. Instead we will impute Census 2000 values to pre-2000 data on a large scale, using the same training data and random forest methods which can be applied on many categories at once, building a sophisticated imputation model for each one.

Dual-coded data sets are the gold standard training sets but some inferences from a Census may apply to a CPS because the Census has near-complete coverage. (Examples are in Meyer, 2010) The NLSY also dual-codes its data into multiple Census schemes and it can be brought to bear in future research to improve sample size and accuracy of imputations of the same kind. We can train our machine learning on a dataset in which the specialist coders from the census have assigned both the Census 1990 and the Census 2000 industries and occupation, to each person.

These prediction models can be trained on several data sets. The prediction variables vary but generally include the individual's age, race, sex, years of formal education, earnings, U.S. state of residence, occupation, and employer's industry in one of the decadal Census classification systems.

5. Our input and output files

Our input CPS files for 2000-2018 have about 50 variables, most of which are listed in Table 1. We append several more and fill in intermediate variables to the 1986-1999 data in the process discussed below. We predict and impute these standardized variables:

- Major and minor occupation categories (2 and 3 digit, in the Census 2010 category system)
- Class of worker
- Census 2000 sector and industry
- NAICS sector and industry (sector being of the type in table 2)

The CPS basic monthly files, combined with the IPUMS-CPS data, have 15.5m observations of individuals. In some versions of imputation we use IPUMS customized variables. For our training data set, we match IPUMS-CPS data for 2000 to 2002, by observation with our CPS basic monthly data set and dual-coded data set. We also add a column with a z-score for the local unemployment rate, because it can help predict time-varying employment.

Table 4. Key variables in CPS 2000-2018 data

The data from 2000-2018 includes 15,545,508 observations. These are the variables used most in this research. We augment the CPS data from before the year 2000 with imputations of variables with these definitions.

Variable	Explanation
HRHHID	Household id (a long number)
Month	1-12
Year	2000-18
MISH	Month in CPS sample, 1-8
State FIPS code	51 categories
Metro area	4 categories
Age	in years
MARST	6 categories
Sex	2 categories
EDUC	Education, coded differently in different time periods
RACE	26 categories
HISPANIC	11 categories
CITIZEN	Citizenship, 5 categories
EMPSTAT	Employment status, 4 categories
PAYABS	3 categories
UHRSWORK 1	usual hours of work per week, e.g. 40
UHRSWORK2	usual hours of work per week on job 1, e.g. 40
AHRSWORK1	average hours of work per week on job 1
AHRSWORK2	average hours of work per week on job 2
AHRSWORKT	Total hours on all jobs, when available
CLASS	Employer type (private, government, nonprofit, etc) in 8 categories, for job 1
CLASS2	Employer type (private, government, nonprofit, etc) in 9 categories, for job 2
WKSTAT	53 categories
NCHILD	Number of children
IND	Industry of job 1, in 275 categories
OCC	Occupation of job 1, in 571 categories
IND2	Industry of job 1, in 275 categories
OCC2	Occupation of job 1, in 571 categories, 543 of which are used
COUNTY	414 categories
Dumex	is 1 if the person has a second job

Table 5: Imputed NAICS Industries and sectors

The source data have a Census industry, in one of several categories. These NAICS-like industry sectors are imputed. The numbered elements have similar definitions in the Current Employment Statistics, Occupational Employment Statistics, and productivity statistics. We show sample size in CPS data from 2000-2018.

NAICS Sector #	Sector title	NAICS industries	Sample size in CPS	Proportion of workforce
10	Natural Resources & Mining	113-115, 21	157,894	1.0%
20	Construction	23	1,130,433	7.3%
31	Durable manufacturing	321, 327, 33	1,036,771	6.7%
32	Nondurable manufacturing	31, 322-326	620,000	4.0%
41	Wholesale trade	42	415,977	2.7%
42	Retail trade	44, 45	1,780,154	11.5%
43	Transportation & warehousing	48, 49	562,994	3.6%
44	Utilities	22	132,920	0.9%
50	Information	51	352,238	2.3%
55	Financial Activities	52, 53	1,042,724	6.7%
60	Professional and Business Services	54-56	1,660,497	10.7%
65	Education and Health Services	61, 62	3361,482	21.6%
70	Leisure and Hospitality Services	71, 72	1,432,440	9.2%
80	Other Services	811-813	674,866	4.3%
FM	Farms		271,018	1.7%
GV	Government		750,017	4.8%
PH	Private households		82,287	0.5%
PO	Post Office		81,695	0.5%
	CPS sample size for 2000-2018		**15,546,407**	

6. Implementation: random forest algorithms in ranger

Programs in the R language, using the package ranger, apply the imputations (Wright and Ziegler, 2017). Ranger uses random forest algorithms modeled to make predictions on training data and applied to a test set. Here the dual-coded data of 2000-2002 is the training data, and the test set is data from 1986-1999.

Random forest algorithms derive from decision tree methods. Decision trees are built in stages, not in one estimate like a regression. In constructing a decision tree, the computer selects sequentially from the independent variables at random to make estimates by regressions or threshold breaks that will help predict the dependent variable. A full decision tree will have

many steps, which can benefit from particular relationships of categories or thresholds between the independent and dependent variables. In our application, for example:

- there are continuous relationships of income and occupation,
- there are discrete relationships between occupations and industries,
- people working in the mining or farming industry will tend not to be in urban areas
- states have concentrations of particular industries,
- and a lawyer cannot have fewer than 16 years of schooling.

Thus many variables help make better predictions, and the predictors have continuous, discrete, and threshold relationships to the predicted variable of occupation and industry.

A random forest is made up of many decision trees which have been seeded to start differently from different independent variables. Each decision tree implicitly makes a prediction from each observation in the data, and these "vote" to select a classification prediction, or are averaged to make a continuous prediction. These decision trees are built based on threshold values and regression results within the training data. The software builds a giant Rube Goldberg machine for each industry or occupation. The arbitrariness of a "random" forest is surprising, but the random forest method has been shown to reduce overfitting relative to other decision tree methods.

Given that the observations to work from have dozens of variables and hundreds of discrete categories, the computerized implementations of random forest allow us to escape from studying each case. However, it is difficult to summarize the prediction, which is a virtual black box. That is an argument against using random forests -- they have arbitrary and hard-to-explain elements. We believe those disadvantages are not significant here if we check the resulting augmented CPS database sufficiently against benchmarks and smoothness criteria. We have judged the random forest to be suitable for our application, partly because the problem has so many inputs and outputs that simpler methods are not capacious enough. There is no possibility that the "true" model is simple. We understand the data well and can judge the variable importance reports.

Ranger uses many predictor variables of various types, which is a kind of flexibility needed for this problem. The resulting decision trees are large, and with many input variables can be slow to run. The computation to get from the 1986-1999 CPS data to the augmented result takes several hours. Ranger can be configured in several ways. For technical details see appendix A.

The diagnostics from the ranger library can list the predictor variables in terms of importance in making the prediction. Each variable's importance is judged by whether the model would predict very differently if that variable were not present.

The program assigns the Census 2000 industries and occupations, and uses them to assign our 61 NAICS industries of interest. We have verified that each category is imputed to some observation.

7. Industry classification example

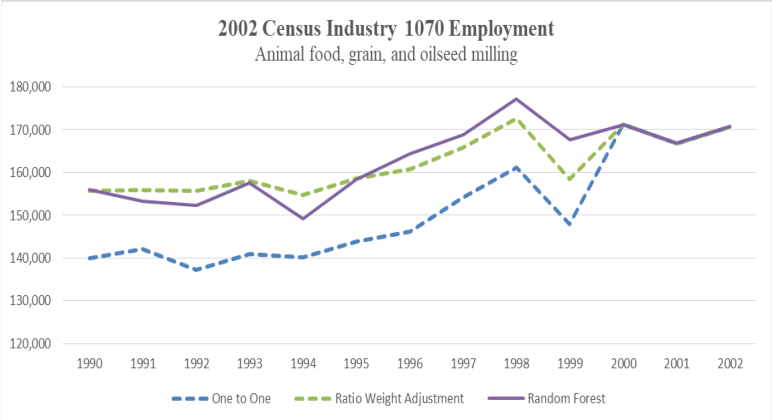
2000 and 2002 Census Industries classifications created a major break in Census and CPS time series. In the CPS dual-coded dataset, 78.5% of observations classified in the 1990 Census

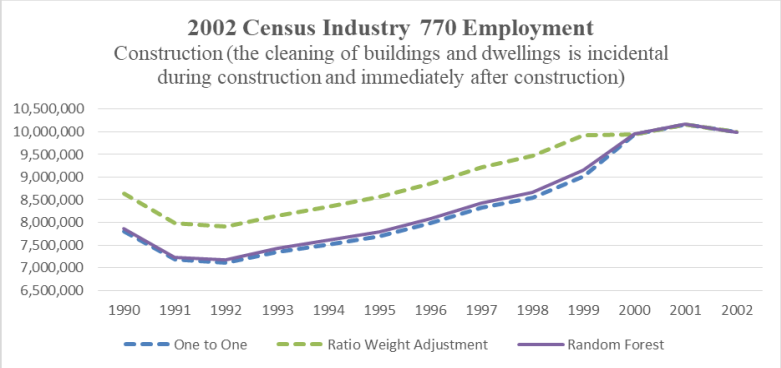
Industry 110 are classified in 2002 Census Industry 1070, both titled “Animal food, grain, and oilseed milling.” If one used a simple cross-walk method, all these would be kept together, though 22.5% belong in a different 2002 industry.

Another method of conversion is to create ratios based on the sum of weights of the dual-coded CPS dataset, split the observations classified in 1990 Census Industries, and adjust the split observations’ weights based on the ratios. This is broadly more accurate than the first method but does not use the microdata from each observation to classify it best by industry.

With a statistical learning approach, each observation is matched to single category in the current classification system, based on its full data, potentially reducing bias and error in estimates for each industry. This method assigns each pre-2002 observation into one 2002 industry. In our tests, random forest algorithms with 500-1000 trees achieve over 90% in-sample accuracy classifying 1990 industry 110 into several 2002 industries, based on about 20 features in the dataset, mainly the worker’s occupation, location, and demographics. With multilayered decision procedures such as random forests, this method can classify more accurately than a single logit could. It is an advance beyond estimating coefficients in one regression. Figure 1 shows the results of the 2002 Census Industry 1070, comparing the random forest approach against the two common alternatives, a one-to-one crosswalk of the kind discussed in section 3, and the ratio adjustment method. The one-to-one approach simply converts 1990 industry 110 to 2002 industry 1070. Figure 2 shows comparisons between conversion methods for 2002 industry 770. For the one-to-one crosswalk this is 1990 industry 60.

The random forest approach has the advantage of using almost all available information, but it does not have the simple interpretation of a single regression or decision tree.





Figures 1 and 2: Employment measures after imputation or Hamilton's method

8. Issues with Census to NAICS conversion post 2000

CPS industries are classified using the Census industry classification system. As many researchers are interested in the NAICS classification system, we studied resolving a common problem in converting Census industries post 2000 to NAICS. Certain industries are underspecified for our purposes, and will not directly convert into a NAICS industry or Sector, e.g. 'not specified manufacturing' must be classified as either durable manufacturing or nondurable manufacturing. The respondent's occupation and other attributes help make this classification. The relevant Census industries for this imputation are 2990, 3990, and 480. The method used is described in Asher et al (2019). We compare estimates from our random forest method to the Hamilton method (figure 3) and share adjustment methods (figure 4). These are methods commonly used to resolve this conversion issue; see Asher et al (2019) and "Largest remainder method" in Wikipedia). We see only modest differences, showing the random forest did not induce noticeable error.

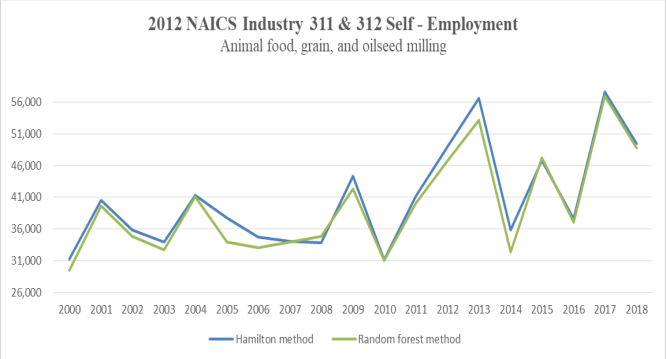


Figure 3: Self-employment measures after imputation or Hamilton's method

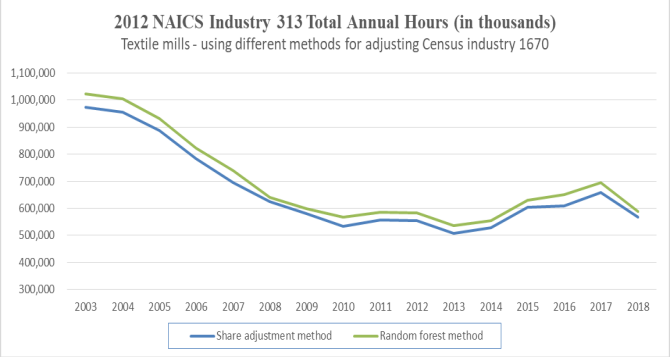


Figure 4: Estimates of hours worked by industry imputation or Hamilton's method

9. Testing augmented files

Applying such algorithms creates an "augmented" CPS data set with predicted industries and occupations for every observation of an employed person from 1986-2018. We can compute employment, self-employment, and hours worked from augmented CPS, just as with the original CPS. This gives us several dimensions to check whether the imputations are plausible. Above we have shown that some imputations are demonstrably acceptable.

Broad tests of the augmented data set are still necessary. One form is benchmarking. The full augmented CPS occupations and industries should match totals in other sources such as the population Census and the QCEW, Furthermore, industries and occupations should evolve slowly in time series of various statistics tracking them from year to year, such as (a) the fraction of the population in the industry or occupation category; (b) their average earnings; and (c) their demographic and geographic distribution. These time series can be tracked and tested for volatility.

If a series were found to be volatile or to miss its benchmarks, the imputations can be adjusted by changing the thresholds that cause a particular category to be assigned. A more advanced method is multiple imputation, perhaps by creating fractional people out of one respondent, and giving different imputed attributes to the fractions.

10. Extensions and further research

If these methods are successful, there are a number of extensions that will further improve imputations of industry and occupation. First, there are other sources of external/dual-coded industry and occupation data:

- The dual-coded 1970-1980 Census sample, called the Treiman data set, discussed in Meyer (2010)
- NLSY (National Longitudinal Survey of Youth) data are dual coded
- Population Censuses can impute some things to the CPS as shown in Meyer (2010)
- The CPS includes the same respondents repeatedly, creating some dual-coded data

Second, there are more data sets beyond the CPS to augment with the same methods. A key example that can help labor composition is the ACS. The Population Censuses can also be likewise augmented, as IPUMS has done.

11. Conclusion

The random forest approach works and gets us key benefits. We are able to assign occupations and industries on a large scale, without analyzing each case ourselves. The input data include individual information on each employed person, dual-coded training data, and big data from other respondents across many years. To our knowledge, this is the first known implementation of a system to impute individual industry and occupation across several Census and CPS data sets based on large scale training microdata. Testing, evaluation, and iteration are necessary before they are usable for production or research work. The resulting augmented data sets are expected to have more accurate long term industry and occupation time series than those now available for social science research.

Appendix A: Implementation and tuning details

This project has more than a thousand lines of source in R so far, processes millions of CPS observations, and draws from substantial training data sets.

There are several R implementations of random forest methods. The ranger implementation seems to suit us. We have not compared it to other implementations.

The model is more accurate when it is allowed to build more decision trees. In some cases we have built only 100 trees so as to stay in memory and run fast enough, but to fully exploit the many predictive categories it appears that a thousand or more are needed. Computer time and memory are limitations for now.

While executing, the software can use more than 5 gigabytes of disk space for the random forest models, and it takes several hours to run. If not carefully configured it would run out of memory or disk space, and sometimes gave errors that would suggest that the problem had been mis-specified when it was simply out of disk space.

There are several tuning parameters, and we are experimenting with the effects of varying them:

- How many decision trees are constructed for each imputed variable. More trees enable more accuracy, but require more time and memory.
- How many branches and variables are used at each branch of the trees. Again more will tend to improve accuracy but requires more resources
- The random seed to start with. This should not have any significant effect, but when the decision trees are too small, it does.
- The proportion used in training versus the test set. We are using 85% in the training set by default.

References

Asher, Kendra; Peter B. Meyer; Jerin Varghese. Improving Census to NAICS industry matches. Poster presented at Data Linkage Day at National Academy of Sciences, Oct. 18, 2019. http://econterms.net/innovation/images/d/d2/Data-Linkage-Day-poster-Oct2019-v8.pdf

Autor, David H.; David Dorn. The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. *American Economic Review*, 2013, 103(5): 1553–1597. http://dx.doi.org/10.1257/aer.103.5.1553

Bureau of Labor Statistics. 1983. Labor Composition and U.S. Productivity Growth, 1948-90. BLS Bulletin 2426.

Dorn, David. Essays on Inequality, Spatial Interaction, and the Demand for Skills. Dissertation, University of St. Gallen no. 3613, Data Appendix, pp. 121-138, 2009.

Meyer, Peter B. Updated unified category system for 1960-2000 Census occupations. Federal Committee on Statistical Methodology conference. 2010.

Meyer, Peter B.; Kendra Asher. Augmenting U.S. Census data on industry and occupation of respondents. 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 600-601. https://ieeexplore.ieee.org/document/8964132, doi: 10.1109/DSAA.2019.00076

Polivka, Anne; Stephen M. Miller. 1995. The CPS After the Redesign: Refocusing the Economic Lens. BLS working paper 269. https://www.bls.gov/osmr/research-papers/1995/pdf/ec950090.pdf

Ruggles, S.; S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas and M. Sobek. IPUMS USA: Version 9.0 [dataset]. Minneapolis: IPUMS, 2019.

Scopp, T. M. The Relationship between the 1990 Census and Census 2000 Industry and Occupation Classification Systems" U.S. Census Bureau Technical Paper #65, 2003.

Wright, M. N.; A. Ziegler. 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77:1-17.