Labor Composition 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
December 31, 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 estimates about the composition of the workforce by industry, researchers want smoother time series for industry and occupation.

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.

We have experimentally applied the same techniques to American Community Survey (ACS) data to further add to the CPS with estimates from this larger but less frequent data source. We compare experimental labor composition indexes from the CPS alone to analogous indexes from the ACS that is benchmarked to selected totals from the CPS.

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

1. Introduction

[1] Views presented by the authors do not represent views of the Bureau of Labor Statistics. We thank Cindy Cunningham, Zhaochen He, Jay Stewart, and participants at the 2020 JSM conference and 2020 GASP workshop for advice.

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 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]

[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

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.

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 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.

---

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.

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: Example rows of an occupation crosswalk

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

Classification of workers into industries and occupations can be ambiguous, and also involves

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

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.

When a new classification is implemented, 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. Two examples from population Censuses are useful to gain some intuition for the peculiarities of these predictions.4

Some Censuses have had "lawyers" and "judges" as separate occupational categories, and others have combined them into one category. Previous research investigated how accurately it was possible to split them apart, when they were pooled in training data from the 1970-1990 population Censuses, and then to split them apart in the 1960 Census. A variety of variables helped make the predictions in a logistic regression. Judges are employed only by governments. The person's age, earnings, and business income helped predict the classification. And a surprise was discovered: if the person had fewer than 16 years of education, it was always a judge. It turned out that while practically all lawyers had gone to law school, a noticeable proportion of judges had not. This unexpected fact showed that perfect predictor.

Likewise, actuaries and statisticians were combined in the 1960 Census but distinguished in later Censuses. Again using the Census variables for industry, person's age, education, and earnings were useful predictors. Apart from these, certain states seemed to have actuaries instead of statisticians: Connecticut, Minnesota, Nebraska, and Wisconsin. It turned out these states had the headquarters of large insurance companies. Thus in several cases of occupations, some unexpected threshold or categories was discovered to be a useful predictor. The random forests method to be discussed can find and use such cases, without a person having to discover them.

[4] Meyer (2010) has the details of the lawyers-and-judges and statisticians-and-actuaries regressions.

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. CPS input files

Three CPS data sets are involved. 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 added a column with a local unemployment rate and a normalized z-score for the local unemployment rate relative to other regions. The unemployment statistics can help predict time-varying employment.

Our CPS data for 2000-2018 has about 50 variables. The key variables used to impute industry are listed in Table A1.

Our input CPS files for 1986-1999 come from IPUMS, and have about 10.4 million observations. For further details see Appendix A. 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 15.5 million observations of individuals. In some versions of imputation we use IPUMS customized variables.

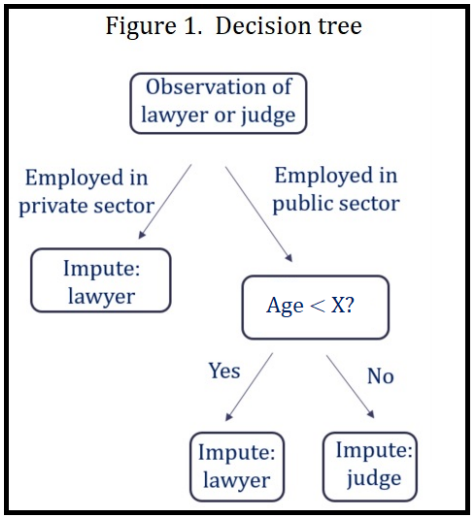
6. Implementation: random forest algorithms in ranger

We make the imputations using R language programs using the ranger package. (Wright and Ziegler, 2017). Ranger uses random forest algorithms modeled to make predictions on training data and applied to a test set. Here the main training data is the dual-coded data of 2000-2002, and the main test set is CPS 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 simple such tree is shown in Figure 1, using only one variable at each step. A full decision tree will have many branches, 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 occupation to likely incomes,
- there are discrete relationships between occupations and likely industries,
- people working in the mining or farming industry will tend not to be in urban areas
- states have concentrations of particular industries,
- and some occupations are held almost entirely by people with certain levels of schooling.

Figure 1. Decision tree



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 classifier is made up of many decision trees which lead to predicted classifications based on the predictor variables. Each branch in each decision tree "splits" the sample of input data on the basis of category predictors or linear combinations of continuous variables, and a threshold value specifying the value of this linear combination to split the data. The split can be represented geometrically as a plane through the data, with values above the plane branching away from the values below.

The variables considered at each branch are randomly selected from the predictor variables, with some constraints set by parameters chosen by the analyst and other structural constraints to prevent the decision trees from being too similar. Each decision tree is constructed by the software package to generate an optimal prediction given the constraints on its own construction and the training data.

In the ranger package, the function ranger() constructs these decision trees given a training data set, a variable to be predicted, and other predictor variables. It produces as many of these trees

as a user has specified. The software thus builds a giant Rube Goldberg machine for each variable to be predicted. The resulting structure, a parameterized model, is stored as a dataframe, an R data set. At a later time the programmer can use the function predict(), giving it this parameterized model and a test data set which has the same variables, and the decision trees each make a prediction and their combined "vote" is the prediction of the model overall.

The resulting parameterized random forest model can make predictions from new test data. When applied to new data, the prediction of the random forest classifier for each observation is defined to be the winning "vote" from the set of the decision trees. A random forest could make predictions for a continuous variable by averaging the predictions of the many trees.

The randomness of the random forest may be surprising, but the method is able to capture many interaction effects among the predictors and has been shown to reduce overfitting relative to other decision tree methods.[5] There are discussions in the literature of when overfitting could still happen, and which kinds of interactions would be missed.

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 C.

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 then assign our 61 NAICS industries of interest. We have verified that each category is imputed to some observation.

8. Post-2000 CPS: Converting ambiguous Census industries to NAICS

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

[5] A neutral net approach could be more accurate but our data and our problem do not seem to call for it, and it would be computationally much more intensive. Neural nets outperform random forests for problems with certain holistic properties, like classifying an image as that of a cat or dog based on the pixels of the image.

problem in converting Census industries post 2000 to NAICS. Certain Census industries are underspecified for our purposes. They will not directly convert into a three-digit NAICS industry or Sector, e.g. 'not specified manufacturing' must be classified as either durable manufacturing or nondurable manufacturing, and then classified into a three-digit NAICS industry before we can use it for our application. Census 2000 industries 480, 1670, 2990, and 3990 are under-specified in this sense.

The respondent's occupation and other attributes help make this classification. The method used is described in Asher et al (2019). We compare estimates from our random forest method to the Hamilton method (figure 2) and share adjustment methods (figure 3). 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 imputation algorithm did not induce noticeable error.

The largest remainder method (called Hamilton's method, in the chart) is what we do in production now.

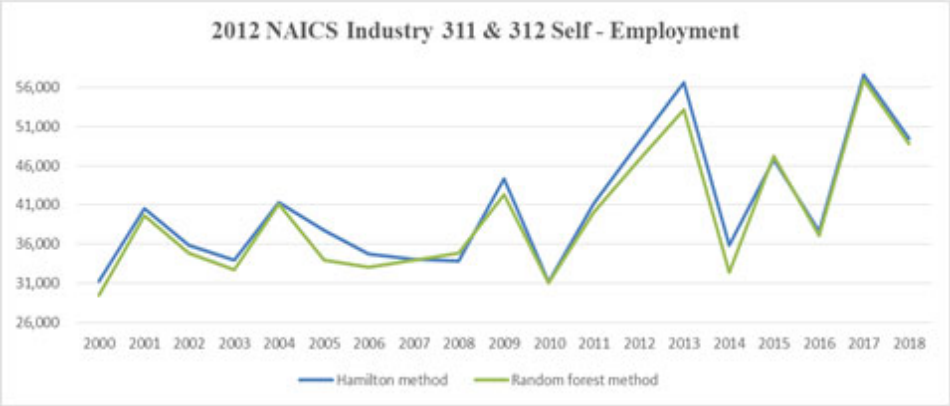


Figure 2: Self-employment measures after imputation or Hamilton's method

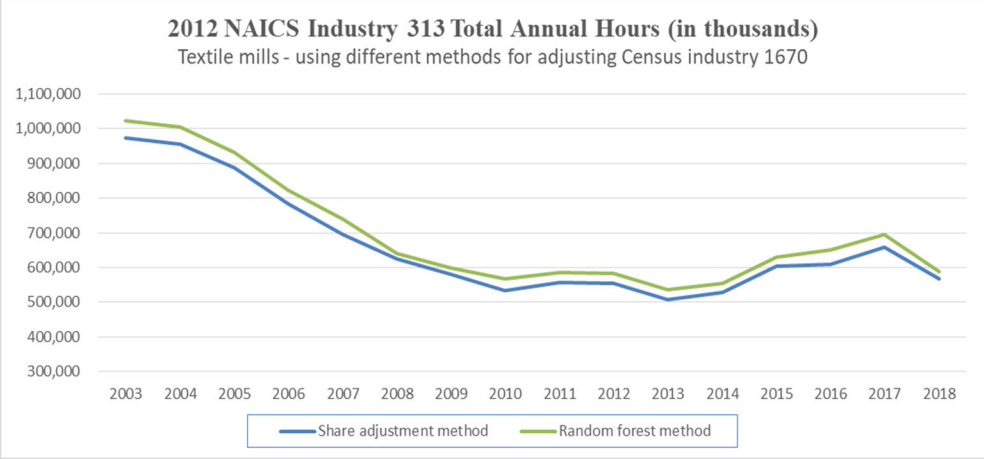


Figure 3: Estimates of hours worked by industry imputation or Hamilton's method

8. Pre-2000 CPS: Industry classification example

More work is necessary on CPS data from before the year 2000 because it uses fundamentally different industry and occupation category systems. This example illustrates an issue with our use of CPS data from before 2000. 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.

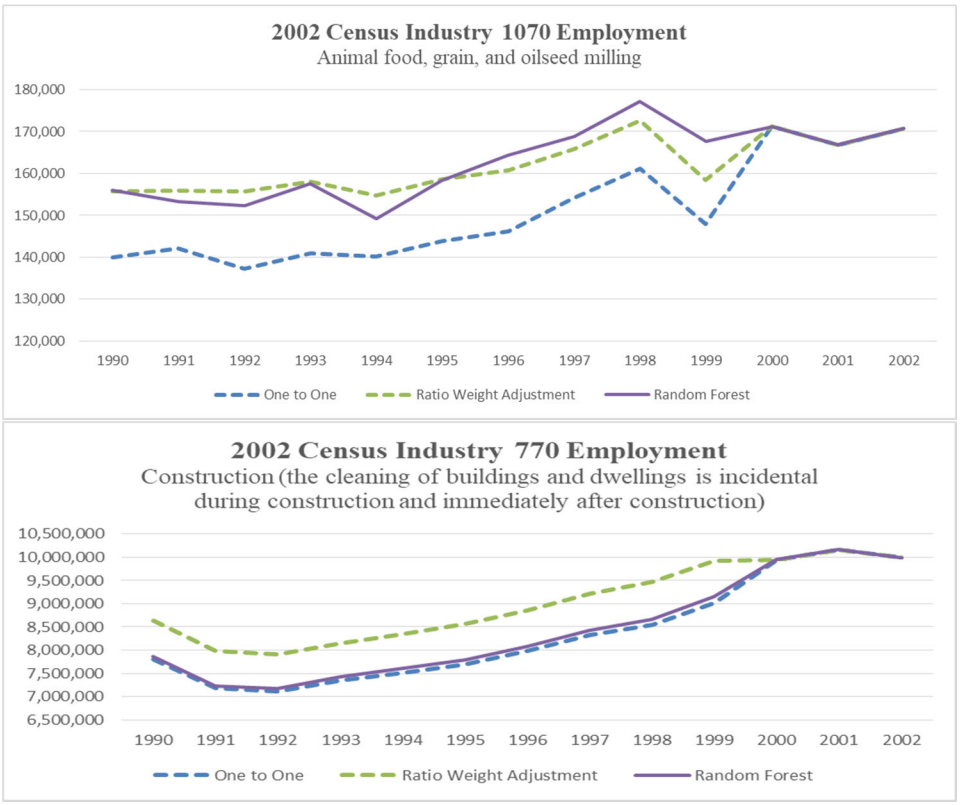
With a statistical learning approach, each observation is matched to single category in the current classification system, based on its full data, potentially reducing bias and error in estimates for each industry. This method assigns each pre-2002 observation into one 2002 industry. In our tests, random 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 4 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 above, and the ratio adjustment method.[6] The one-to-one approach simply converts 1990 industry 110 to 2002 industry 1070. Figure 5 shows comparisons between conversion methods for 2002 industry 770. For the one-to-one crosswalk this is 1990 industry 60.

---

[6] In Figs 2-5 we compute employment estimates directly from the CPS data. The purpose of the chart is to show variations coming from our methodology of classifying observations by industry. Each employed respondent has a weight, and combining the weights for a group of employed persons gives an estimate of how many there were in the U.S. population. Discussant Jon Samuels has pointed out that the upward trend in employment observed in Figure 4 is contrary to the downward trend in the 1990s observed in the CES measures for NAICS 3111 and 3112 which should correspond to Census 1070. Weighting, self-employment, and sampling and definitional differences between CPS and CES may explain a difference. CPS aggregates will not perfectly match CES (Current Employment Statistics) aggregates, which come from establishments. CPS data comes from households, not establishments, and the weights are constructed by demographics, not by establishment. Therefore, normally one assumes that CES data are more precise for establishment industry estimates. CPS establishment industry estimates are normally benchmarked to the CES, QCEW, or other industry-focused data. For our CPS employment count comparison in Figure 4, we counted people who were recorded as employed in the previous week. Challenging cases include (a) those classified as employed by government, but in a private industry too, (b) the self-employed, or (c) those with a second job in this industry. For this particular test we are only trying to see if we get similar numbers from CPS's with different imputations, before benchmarking.

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.

The random forest line in purple is (plausibly) the smoothest and therefore probably the most realistic. It would be thinkable to construct estimates by month and test for volatility.



Figures 4 and 5: Employment measures after 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, such as those in the figures above, to check whether the imputations are plausible.

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. Extension to ACS and labor composition indexes

Labor composition indexes summarize changes in the age and education levels of the workforce in each industry. We use an established method that uses earnings in groups distinguished not only by industry, but also age, sex, and education levels. The workforce proportion, wage share, and earnings share of each cell defined by these variables is combined in a Tornqvist index. Our office has a normal procedure to produce these from the original CPS data with a crosswalk of imputations, and we extended this procedure to test it with the augmented CPS discussed and separately with the augmented ACS discussed below.

The ACS (American Community Survey) data has similar variables to CPS, including industry and occupation, and has a much larger sample size allowing us to more accurately measure finer detail. However, ACS data is gathered largely by mail, and is therefore less precisely managed than CPS data; it appears to have more errors and internal incompatibilities. See Appendix B for some comparisons of these data sources. For our intended use, CPS weighting methods are more accurate, and therefore CPS totals using those weights for values like average hours worked and average wage are more precise. Our goal was to create better microdata to compute industry estimates like labor composition. To do so we created a combination of ACS microdata summing to CPS economy-wide totals using machine learning.

First, we obtained ACS data from IPUMS, and then augmented it with select variables as with the 2000-2018 CPS (discussed in section 7) to allow for better algorithms in our machine learning programs. Our IPUMS data with the altered augmented variables from 2003-2018 has 20.9 million observations of 85 variables. We next adjusted variable definitions, when necessary to account for longitudinal changes in definitions, other than those of industry and occupation. Then we were able to create algorithms which converted observations from split or unspecified Census Industry to NAICS. We followed a similar course to adjust CPS data for the years after 2000.

Our next stage was raking the ACS data to CPS totals for select variables. With a labor composition estimate in mind, we raked the ACS weights, that we use to create total hours and average wage, to CPS totals by two-digit sector, age group, education, and sex. We used three-year weighted moving averages from the CPS. Our reason for using the three-year weighted moving average was to account for CPS's low sample size in certain cells. We varied the weights for the moving averages during testing. For example, in one test we used $year_{t-2}$=20%, $year_{t-1}$=30%, and $year_t$=50%. Controlling to CPS totals also helped us account for a main difference between ACS and CPS data. One main advantage of CPS data, aside from the weights being more accurate for our uses, is the inclusion of second job data. By matching CPS sector, age

group, education, and sex totals, with this distinction added, we are indirectly adding the second job distinction to our ACS measure.

After raking our ACS data to match CPS totals, we construct a labor composition measure. For comparison, we also created this labor composition measure by using data which rakes the current year CPS weights to a matching CPS three-year moving average. This differs from our standard labor composition methodology. For all years prior to 2003 we used CPS based raked data—which also incorporates our machine learning algorithm to convert pre-2000 industries—for both the ACS and CPS LC measures. Thus, the index's growth rate should be identical between the two measures in these years. (Note: as there are large methodology differences, neither measure will match BLS published labor composition measure). The variance of growth for our ACS based measure was slightly smaller overall than the CPS based measure, although for many industries there was not a significant difference.

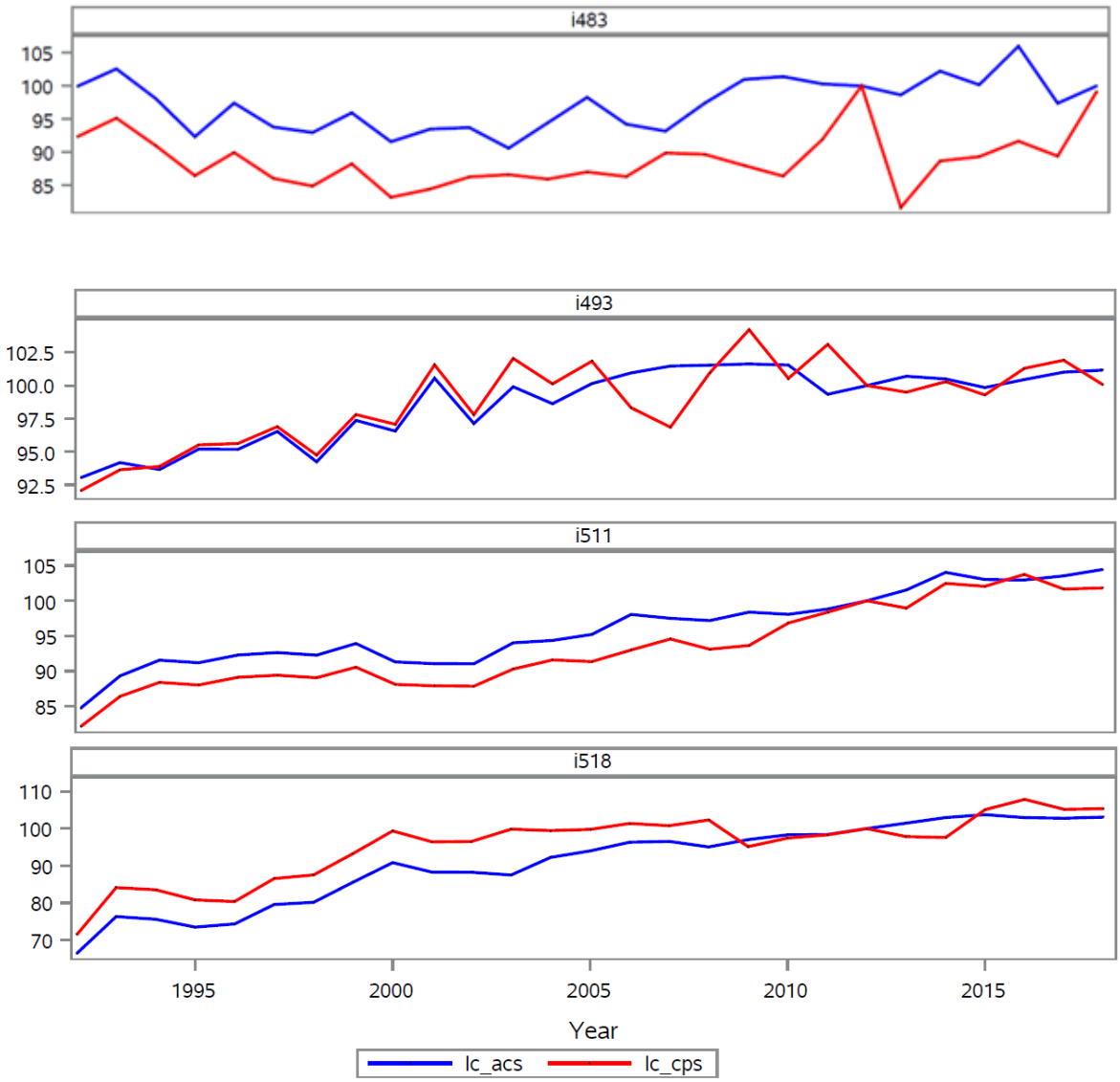


Figure 6: Experimental labor composition indexes from ACS (in blue) and CPS (in red)

In the charts in Figure 6, the data before 2003 are all from CPS so the underlying data has an identical growth rate. Both series are indexed to 100 for the year 2012.

Both series are controlled, or benchmarked, to CPS totals for two-digit NAICS sector, sex, and age and education groups. We used a RAS procedure. The differences between the series are due to variation at the 3-digit industry level in the samples. Labor composition incorporates wages, hours and weights for each.

The ACS weights each year are scaled to meet hour weights and wage weights in three-year economy-wide averages from the CPS. The weighted average of the CPS years used in the charts above weight $year_{t-2}$ by 20%, $year_{t-1}$ by 30%, and the reference $year_t$ by 50%. Both CPS and ACS are weighted that way.

The ACS indexes generally looks better here, in that they exhibit less variance. The CPS for 483 in 2012 happens to be volatile – high – in 2012. This is why the two lines are seem to be apart. But their growth rates, which are used in productivity estimates, are overall close together.

Thus we find that generally the ACS can help create labor composition indexes for smaller industries, once they are benchmarked (controlled) to the 2-digit industry from the CPS. We expect the CPS to be better for 2-digit data because the data are of higher quality -- more carefully scrutinized by interviews.

11. Further applications

If these methods for imputing industry and occupation are successful, there are a number of improvements and extensions that follow. First, there are other sources of external/dual-coded industry and occupation data to treat as input for augmenting the CPS:

- The dual-coded 1970-1980 Census sample, called the Treiman data set, discussed in Meyer (2010) can be applied to extend imputations to CPS back to the 1970s.
- NLSY (National Longitudinal Survey of Youth) data are dual coded.
- Population Censuses can impute some values to the CPS as shown in Meyer (2010)
- The CPS includes the same respondents over the months, creating some dual-coded data implicitly.

Then, there are more data sets beyond the CPS and ACS to augment using the same methods. The Population Censuses through 2000 can also be augmented in a similar way as illustrated by Meyer (2010). IPUMS has already added many standardized variables for each of these data sets, though not, so far as we know, by using detailed individual-specific imputations as is done here.

12. 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: CPS and ACS input files

We have four basic data sets:

- The CPS from 1986 to 1999, including any variables we can get from IPUMS
- The CPS from 2000 to 2018, including any variables we can get from IPUMS. This is distinguished from the earlier data because it uses the Census 2000 classification systems
- The ACS from 2003 to 2018, which is used and discussed in section 10
- The "bridge" data set from 2000-2002 which gives the training to the machine learning model used to fill in most of the imputations discussed.

The "bridge" data set is distinctive in this paper. It combines the dual-coded data set created by the CPS program originally with the basic CPS for 2000-2002 and some variables from IPUMS for the same months. For most observations, the match was by the variable `UH_UNIQHH_1 LINENO`. The exceptions are associated with an apparent problem in the source data. In the bridge data set alone, a particular set of months in 2000 and 2001 have an oversample, and these could not be matched the CPS data to the IPUMS-CPS. In the IPUMS data set they include an oversample that isn't in the (NBER) version of the CPS.[7] We wrote a SAS program which matched these observations to one another based on several variables, and were able to make use of the extra observations. We did not use the March CPS ASEC (the March CPS) nor the ORG (outgoing rotation group) detailed data.

In the next tables we show summary statistics and variable attributes in the other data.

[7] The issue was discussed on an IPUMS forum: https://forum.ipums.org/t/cps-sample-size-april-june-2001-discrepancy/2537

Table A1. 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 listed in Table A2. Each observation characterizes an individual with a job, in a household whose attributes are known.

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	Whether employee would be paid if absent from work, 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	US counties, 414 categories
Dumex	is 1 if the person has a second job

We augment the 1986-99 CPS data with imputations of variables, mostly from IPUMS. The UH_ variables are in IPUMS only. "UH" means "un-harmonized." IPUMS may have made some adjustments to the codes which are convenient for our imputations.

The two unemployment measures are derived from the state-level unemployment rates for the respondent's state and month. This public data comes from the BLS Local Area Unemployment Statistics (LAUS) program. The z-score is normalized relative to the population of states at that same time period, from each month.

Table A2. Key variables from IPUMS-CPS or BLS-LAUS

Variable	Explanation
UH_HOURSX_1	Hours worked
UH_HOURS_1	Hours worked
UH_IND2JB_1	Industry of 2nd job, from IPUMS, in the native classification
UH_CLASS2JB_1	The class of the 2nd job, from IPUMS (9 categories)
UH_OCC2JB_1	The original (native) occupation of the 2nd job, from IPUMS (472)
UH_PAYABS_1	Yes/no/other if paid for absences on first job, from IPUMS[8]
UH_PAYABS_2	Yes/no/other if paid for absences on first job, from IPUMS
z_unemp	Normalized z-score of state area unemployment, from BLS/LAUS
Unemp_diff12	Unemployment rate in current month minus the unemployment rate a year earlier for this person's U.S. state, from BLS/LAUS

[8] Several variables including the PAYABS variables are defined differently between the decades.

Table A3 shows how many observations there are in the original CPS and ACS data to which we make imputations, and how many original industry and occupation categories were assigned before our imputations. Our ACS data starts in 2003. Some of the occupation and industry categories are not relevant for labor composition – government or private household work – and we drop these, and do not use any imputations for them. Some other industry detail disappears because we are controlling to NAICS-industry totals, which is a higher level of aggregation.

Table A3: Numbers of data observations and occupation and industry categories each year

Reference year	CPS sample size	ACS sample size	CPS industry categories	ACS industry categories	CPS occupation categories	ACS occupation categories
1986	782,588		228		389	
1987	784,020		226		388	
1988	750,602		227		391	
1989	766,641		228		389	
1990	800,938		228		393	
1991	791,083		229		394	
1992	780,336		236		392	
1993	766,466		236		456	
1994	758,453		236		456	
1995	750,587		236		453	
1996	668,084		236		451	
1997	674,389		237		455	
1998	680,176		236		449	
1999	683,621		237		455	
2000	760,824		259		503	
2001	814,552		259		502	
2002	886,394		259		503	
2003	873,198	557,411	264	267	503	477
2004	858,156	557,421	264	267	502	477
2005	856,847	1,336,487	264	266	501	469
2006	853,162	1,390,898	264	266	501	469
2007	842,879	1,399,724	264	266	502	469
2008	839,147	1,434,979	264	265	503	469
2009	840,396	1,385,319	263	265	503	469
2010	833,290	1,373,821	263	265	502	491
2011	816,642	1,362,403	263	265	533	491
2012	809,396	1,380,083	263	264	532	478
2013	799,604	1,411,741	260	264	484	478
2014	797,645	1,422,106	260	264	484	478
2015	779,344	1,442,336	260	264	484	478
2016	777,533	1,459,560	260	264	484	478
2017	765,461	1,488,986	260	264	484	478
2018	741,038	1,509,092	260	267	484	529
Total	25,983,492	20,912,367				

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. In Table A4 we show the sample size in CPS data from 2000-2018, of those observations of persons with jobs. These are the sectors whose totals are RAS'd – wage weights and hours weights, which are derivatives of the CPS or ACS weights. For applications other than labor composition, one might want different weighting scaling, e.g. to include occupations.

Table A4: Imputed NAICS Industries and sectors

NAICS Sector #	Sector title	NAICS industries	Sample size in 2000-2018 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**	

Appendix B: Comparisons between ACS and CPS

The table shows differences between the ACS, the basic CPS variables, and the data available from the CPS outgoing rotation groups which are used in constructing labor composition.

Table B1. Comparisons between ACS and CPS and CPS ORG

Topic	CPS ORG attributes	CPS non-ORG-month attributes	ACS attributes	Our steps
Workforce totals	Random sample weighted to match almost whole US workforce	Continuation of CPS ORG households	Less accurately measured than in CPS	RAS using totals from CPS, forcing ACS to match by sector, age group, education group, and sex
Not specified industries Census 2990, 3990, 480	Missing information	Missing information	Missing information	Impute sector (if needed) and industry within both CPS and ACS based on similar respondents
Actual hours last week	has respondent data	has respondent data	Missing information	Impute from CPS to ACS though RAS of ACS usual hours
Hourly wages	has respondent data	Missing information	Missing information	Impute from ACS using other ACS variables, then adjusted by CPS marginal totals.
2nd job if any	Presence recorded in CPS in and since 1994	Missing information	Missing information	No action
Self-employment income	Missing information	Missing information	Recorded	No action
Year coverage	Back to 1986, with methods changes in 1994	Back to 1986, with methods changes in 1994	Back to 2003	We did not use earlier ACS because of definitional changes. We could go further back in either database.

Relevant differences between CPS and ACS data:[9]

- ACS does not have actual hours worked last week. By controlling relevant ACS data to CPS, we are imputing a figure for this to the ACS.
- Estimates of hourly wages are imputed from CPS ORG data to the ACS, again by benchmarking.
- ACS has no relevant information on second jobs. By including second jobs in the CPS totals, we are increasing the weights for hours and scale the ACS up to include the 2nd jobs implicitly. Wages on 2nd jobs are not currently included in either data set so 2nd jobs won't have any effect on wage weights. The difference between the presence or absence of 2nd jobs is a problem for labor composition construction in principle, because we missing some of the changes in people and wages in that industry.

[9] Drawn partly from Census Fact Sheet, Webster (2007), and Kromer and Howard (2011). See also IPUMS documentation.

- Likewise we are missing wage data for the self-employed. ACS has some wage data for the self-employed, but the CPS does not. We do have annual income for the self-employed in the CPS. Self-employment income could be thought to include ownership/proprietorship income. We could make imputations in principle.
- ACS is collected mainly by mail; CPS is collected more precisely by interviewers. CPS interviewers can do follow-up so attributes are more accurately measured.
- Time frames of the respondent data are different. CPS variables refers to the current month and ACS to the past year. (Occasionally we adjust for this.)
- ORG data is used in labor composition – for wages, which are not available in the other months of the CPS. And ACS has wages which we must adjust to CPS's definition.
- INDNAICS in the ACS is a NAICS industry assigned to each observation based on a Census industry crosswalk. It does not disambiguate some cases we need to separate out with machine learning.

Appendix C: Implementation and tuning details for ranger

There are several R implementations of random forest methods. The ranger implementation seems to suit our application (Wright and Ziegler, 2017). We have not compared it to other implementations.

This project has more than a thousand lines of source in R so far, processes millions of CPS or ACS observations, and draws from substantial training data sets. 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 runs out of memory or disk space, and sometimes gave errors that would suggest that the problem had been misspecified when it was simply out of disk space.

There are several tuning parameters which we are experimenting with:

- How many decision trees are constructed for each imputed variable. More trees enable more accuracy, but require more time and memory. Computer time and memory are limitations for now.
- How many branches and variables are used at each branch of the trees. More improve accuracy but requires more computer time and memory.
- 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 of the bridge data set used for training versus measuring accuracy. We estimate accuracy for each imputation by training on 85% of the bridge data set, and testing accuracy on the remaining out-of-sample 15%. The resulting model is then applied to the years for which accuracy is unmeasurable. In a later stage, when we have tuned the system for accuracy, the whole bridge data set can be used for training.

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.

IPUMS. Documentation on INDNAICS. https://usa.ipums.org/usa/volii/indnaics18.shtml

Izrael, David; David C. Hoaglin; and Michael P. Battaglia. 2000. A SAS macro for balancing a weighted sample. Paper 258-25. Abt Associates Inc., Cambridge, MA. https://support.sas.com/resources/papers/proceedings/proceedings/sugi25/25/st/25p258.pdf

Kromer, Braedyn K.; David J. Howard. 2011. Comparison of ACS and CPS data on employment status. SEHSD-WP2011-31. Social, Economic, and Housing Statistics Division, U.S. Census Bureau. https://www.census.gov/library/working-papers/2011/demo/SEHSD-WP2011-31.html

Meyer, Peter B., and Anastasiya Osborne. 2005. Proposed category system for 1960-2000 Census occupations. BLS Working Paper 383. http://www.bls.gov/ore/abstract/ec/ec050090.htm

Meyer, Peter B. 2010. Updated unified category system for 1960-2000 Census occupations. Federal Committee on Statistical Methodology conference. https://nces.ed.gov/FCSM/pdf/2009FCSM_Meyer_IV-B.pdf

Meyer, Peter B.; and 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; and Stephen M. Miller. 1995. The CPS after the redesign. 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. 2019. IPUMS USA: Version 9.0 [dataset]. Minneapolis: IPUMS, 2019.

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

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