Census–NAICS 2012 Industry Code Crosswalk
Documentation File

Evan J. Soltas*
esoltas@mit.edu

Version 1.0
Last updated: 29 May 2018

Description

The enclosed files are crosswalks between the 2012 editions of the Census industry and North American Industry Classification System (NAICS) codes, from 2- to 6-digit NAICS codes. Two files are enclosed:

ind_to_naics.csv	Crosswalk to project data measured in Census industry codes to a NAICS code level
naics_to_ind.csv	Crosswalk to project data measured at a NAICS code level to Census industry codes

Instructions

Using standard statistical software (e.g., R, Stata, Python), users should merge their data by one-to-one or many-to-one merge into the appropriate file. The merge keys are respectively *ind* for the Census industry to NAICS code crosswalk and *naics* for the NAICS to Census industry code crosswalk. It is critical that users keep only the appropriate digit-level subset of the NAICS codes and not merge simultaneously into multiple levels of the NAICS codes.

Motivation

U.S. government data use two related, but nonetheless distinct, sets of industry classification codes. For example, in Business Register products such as County Business Patterns, industries are coded according to NAICS codes, whereas in American Community Survey and Current Population Survey microdata code, workers' industries are coded according to Census industry classification codes. The discrepancies between Census industry and NAICS codes present a considerable obstacle to economists who wish to link data on firm and worker outcomes.

To the best of the author's knowledge, no official or reliable public microdata has been double-coded with Census and NAICS industry codes since their 2012 revisions. The only official

* I release the crosswalk files and documentation under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. Noncommercial users are therefore free to share and adapt this work, provided that they credit this work and that they share their work under the same license. Commercial uses of the crosswalk files and documentation are forbidden. For more, see https://creativecommons.org/licenses/by-nc/4.0/.

resource is a "crosswalk information" file provided by the U.S. Bureau of Labor Statistics (BLS) for use in mapping from Census to (but not from) NAICS industry codes.[1]

The BLS file alone is unfortunately inadequate for use in economic research. Most importantly, it lacks allocation factors when Census industry codes do not map into a unique NAICS industry code. In these cases, the BLS file simply lists NAICS codes that correspond at least in part to a given Census industry code. In addition, the BLS file is critically incomplete in some cases as a mapping from Census industry to NAICS codes. There are several Census industries that are mapped to NAICS categories such as "3118 exc. 311811," leaving it to the user to determine the NAICS codes in 3118 that are not 311811, or yet more cryptically, categories such as "Part of 331 & 332." Furthermore, the BLS file is inconsistent as to the digit level of the NAICS codes to which it maps. Without further researcher effort, the BLS file is not usable as a crosswalk.

Construction

I constructed this crosswalk relying heavily on the BLS file's judgment as to the mapping of Census industry to NAICS codes. My procedure was as follows:

- Where the BLS file prescribes that a given Census industry code maps to a unique NAICS code, I preserved this mapping.

- Where the BLS file provides multiple NAICS codes to which a given Census industry code corresponds, I did not deviate from these decisions. My file, however, provides allocation factors using employment counts from County Business Patterns (CBP) data to guide the mapping from Census industry codes to NAICS codes.

- Where the BLS file was ambiguous or incomplete as to the mapping from Census industry to NAICS codes, I reviewed both Census industry and NAICS code documentation and matched the Census industry code to one or multiple NAICS codes. In the case of multiple NAICS codes, I allocated across NAICS codes according to CBP employment. In general, I deferred to higher-level NAICS categories and relied upon the allocation step. For example, in almost all situations in which the BLS file indicated a Census industry code mapped to NAICS code such as "Part of 331 & 332," I allocated the Census industry code to all of NAICS codes 331 and 332 proportional to CBP employment.

- To map Census industry codes to NAICS code digit levels below the level specified in the BLS file, I allocated according to CBP employment counts. This step was critical in simply constructing a mapping of Census industry codes to a constant-digit level of NAICS codes.

- To map Census industry codes to NAICS code digit levels above the level specified in the BLS file, I exploited the logical nested structure of the NAICS codes and summed upwards. For example, if 70 percent of a given Census code is allocated to NAICS code 54171 ("Research and development in the physical, engineering, and life sciences") and 30 percent to 54172 ("Research and development in the social sciences and humanities"), then 100 percent of the same Census code must be allocated to NAICS code 5417. By

[1] "2012 Census Industry Classification," https://www.bls.gov/cps/cpsoccind.htm.

implication, 100 percent of the same Census code must be allocated to NAICS codes 541 and 54.

- This procedure yielded allocation factors that map from each Census industry code to NAICS codes of all digit levels according to CBP employment. To construct the inverse mapping from NAICS codes of any constant digit level to Census industry codes, I used my Census-to-NAICS crosswalk to identify, for every NAICS code, the Census industry codes that should receive a positive allocation factor. I then used U.S. Current Population Survey data on weighted employment counts by Census industry to construct the factors.

Data

I used several public-use datasets dated approximately 2016 in the construction of this crosswalk. Researchers should exercise caution in using this Census–NAICS crosswalk for vintages of data to the extent they are distant in time from 2016. These datasets are:

- 2015 and 2016 County Business Patterns: National data file downloaded from American FactFinder (factfinder.census.gov) for all NAICS category levels. I then took the average employment count across these CBP years.

- U.S. Current Population Survey data, January 2015 to December 2017: Public use microdata file from IPUMS, Minnesota Population Center, University of Minnesota (cps.ipums.org). I pooled both the basic CPS and ASEC files and collapsed sampling weights to by Census industry code.[2]

Performance of Crosswalk

Viewing the crosswalk as data compression, I assess the performance of the crosswalk in terms of "lossyness" in mapping between Census industry and NAICS codes. If the same concept is measured independently and at high precision with respect to Census industry and NAICS codes, then a crosswalk is said to be a lossless compression if the correlation between crosswalked and actual values is unity. The lower the correlation, the lossier the compression. The slope and intercept of the relationship between crosswalked and actual values are also objects of interest.

I provide results for mapping employment shares, which is of interest to economists and I expect to provide a lower bound on the performance of the crosswalk in standard uses in economics. In the performance test, I crosswalked CBP employment shares into Census industry codes and CPS employment shares into NAICS industry codes. I believe this to be a relatively demanding test of the crosswalk, as there is no ex-ante reason to believe that a mapping error will lead to an industry of approximately the same employment share as the correct mapping. By comparison, the performance of the crosswalk for values such as average wages is likely to be stronger, as even when the mapping is incorrect, it seems likely that the correct and misallocated industry have similar average wages.

[2] One exception: There are no observations of Census industry code 8890 ("Footwear and leather goods repair" in the CPS files I used. I assumed that it had the same number of observations as the next-smallest Census industry code, 7170 ("Video tape and disk rental").

In Table 1, I regress crosswalked on actual log employment shares. I find that crosswalk performance is generally strong, as indicated by the high R^2 statistics throughout the results. The crosswalk performance remains notably strong when mapping into the relatively high-detail 6-digit NAICS codes. Performance is comparable in either direction, NAICS to Census or Census to NAICS. However, the scaling is consistently wrong in mapping from Census industry to NAICS code employment shares. In particular, crosswalked employment shares underestimate the share of large NAICS codes and overestimate the size of small NAICS codes, consistent with attenuation towards the mean industry share due to measurement error. This attenuation bias is notably absent in the NAICS-to-Census direction.

Conclusion

Pursuant to the license agreement under which this crosswalk is released, I ask that crosswalk users cite this document. I welcome users to contact me with questions, use cases, concerns, suggestions, errors, etc. I hope to refine this crosswalk in future versions in the coming months.

Table 1: Diagnostic Table of Log Crosswalked and Actual Employment Shares

	NAICS → Census			Census → NAICS		
NAICS Digits	Slope	Intercept	R^2	Slope	Intercept	R^2
2	1.017 (0.018)	0.103 (0.124)	0.924	0.814*** (0.049)	-0.709*** (0.233)	0.603
3	1.080** (0.032)	0.381* (0.202)	0.834	0.880** (0.051)	-0.621** (0.281)	0.849
4	1.023 (0.037)	0.025 (0.228)	0.765	0.872*** (0.031)	-0.837*** (0.208)	0.834
5	0.973 (0.035)	-0.226 (0.215)	0.764	0.898*** (0.023)	-0.797*** (0.175)	0.812
6	0.935** (0.031)	-0.417 (0.195)	0.775	0.918*** (0.016)	-0.674*** (0.130)	0.850

Notes: This table reports regression results of crosswalked employment shares on actual employment shares. In the left panel, the dependent variable is the CBP employment share, measured originally in NAICS codes and crosswalked into Census industry codes, on the CPS employment share, measured originally in Census industry codes. In the right panel, the dependent variable is the CPS employment share, measured originally in Census industry codes and crosswalked into NAICS codes, on the CBP employment share, measured originally in NAICS codes. All regressions are unweighted. Standard errors are heteroscedasticity-robust. For the slope coefficient, p-values are reported with respect to a null hypothesis of unit slope, whereas for the intercept, p-values are reported with respect to a null hypothesis of zero intercept. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.