Augmenting U.S. Census data on industry and occupation of respondents

Peter B. Meyer
Office of Productivity and Technology
U.S. Bureau of Labor Statistics
Washington, DC, USA
meyer.peter@bls.gov

Kendra Asher
Office of Productivity and Technology
U.S. Bureau of Labor Statistics
Washington, DC, USA
hathaway.kendra@bls.gov

***Abstract*—The U.S. Census Bureau classifies survey respondents into hundreds of detailed industry and occupation categories. The classification systems change periodically, creating breaks in time series. Standard crosswalks and unified category systems bridge the periods but these often leave sparse or empty cells, or induce sharp changes in time series. We propose a methodology to predict standardized industry, occupation, and related variables for each employed respondent in the public use samples from recent Censuses of Population and CPS data. Unlike earlier approaches, predictions draw from micro data on each individual and large training data sets. Tests of the resulting "augmented" data sets can evaluate their consistency with known trends, smoothness criteria, and benchmarks.**

Keywords—Census, CPS, ACS, prediction, imputation, occupation, industry, population, employment

I. INTRODUCTION

Several survey data sets use the same industry and occupation classifications in the U.S. Censuses of Population, notably the monthly Current Population Survey (CPS) and the annual American Community Survey (ACS). Classification changes each decade, or more often, cause breaks in the time series used for official statistics and research studies. Longer time series can be constructed with coarse standardized categories. This is common in social science [1][5]. There are tradeoffs between precision and accuracy: Narrow categories may be necessary for a research question, but have more classification errors than broad ones, and are more likely to leave some industry-year cells empty. Here we show how predictions from based on rarely-used training microdata and economy-wide benchmarks can more accurately classify observations by industry and occupations and aid studies using those categories.

The relevant 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. Electrical engineers, for example, were category 12 in the 1970 Census, 55 in 1980 and 1990, and 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. This project can smooth and help analyze the time series around the definitional classification breaks.

II. TRAINING DATA

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.

"Dual-coded" data sets are monthly CPS samples in which the industry and occupation have been coded into two different Census systems, by the same specialists who would normally classify such observations [6]. At the time the 1980 Census was conducted, a sample of 122,000 observations was dual-coded into both 1970 and 1980 classifications.

Another dual-coded data set 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 [6]. It has 2.4 million observations, with overlap as households were surveyed repeatedly.

Dual-coded data sets are the gold standard training sets, but as we illustrate in the next section, a Census may be a training set for another Census.

III. OCCUPATION CLASSIFICATION EXAMPLES

In the 1970-1990 Censuses, lawyers and judges were distinct occupation categories, but in the 1960 Census they were all in one category. The standard practice when reclassifying them into the 1990 system is to label them all lawyers [5]. We try to infer which of them were judges, training on a pooled set of lawyers and judges in 1970-1990 Censuses. In the training set, all judges were employed by governments; anyone with less than 16 years of education was a judge; judges tended to be older, to work for state governments, to have a high salary, and to have little business income. A logit regression estimated coefficients on these predictors. The resulting decision rule to classify judges and lawyers was 83% accurate on 1970 Census data, a training set [4]. Applying the rule to 1960 Census data, we can adjust the threshold of prediction to fit the number of judges expected. This reclassification improves the accuracy of classification in the lawyers group, extends our time series of judges, and makes better derived estimates about them possible.

Using the dual-coded data sets, we identified predictors of which 1980 category fit persons in a 1970 category, achieving training accuracy of 73% in dividing "personnel and labor relations workers" into their 1980 category successors [4]. Table I shows examples of prediction of the 1990 occupation when the 2000 occupation is known, based on a third data set, the dual coded 1990-2000 data set [4]. Thus we can predict occupation from several training data sets and estimate error.

TABLE I. 1990 OCCUPATION PREDICTORS GIVEN 2000 OCCUPATION

2000 category	1990 category	Predictors	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	

Each mapping is constructed from a study of selected occupation categories. If the same process were applied to all occupations, results might poorly match economy-wide data for a variety of reasons. Imputation of industry and economy-wide benchmarking can help, as discussed below.

IV. Industry Classification example

2000 and 2002 Census Industries classifications created a major break in Census and CPS time series. Census Industries classifications were converted to the 1997 North American Industrial Classification System (NAICS). 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, we classify better than with logits and advance beyond estimating coefficients in one regression as in section III.

V. Testing and Benchmarking augmented files

Applying such algorithms creates an "augmented" Census or CPS data set with predicted industries and occupations. Shifts in population among industry and occupation categories should be slow, and in each augmented data set this is testable. For each industry and occupation category, we expect to see smooth time series of (a) the fraction of the population in it; (b) its average earnings and earnings variance; and (c) its age, sex, race, and location distribution. Each subgroup can be tested for this.

The augmented data set can be also tested against population benchmarks drawn from the decennial Census, the QCEW and LEHD databases, and the national accounts (NIPAs) as in [2]. One reason the augmented data would tend not to match population benchmarks is that the imputation procedures tend to be biased toward conventional outcomes especially when predictions of industry and occupation are done sequentially as here. If there are, for example, 22-year-old chief executives or an elderly farmhand with advanced degrees in the data, prediction-based reclassifications will often impute more conventional roles to them, thus reducing the variance in distributions too far. Tests of the augmented data set will help diagnose in which subgroups this problem arises.

We can iterate to reimpute for individuals to improve the results of such macro tests. One technique is to incorporate a random error into probabilistic assignments to reinflate the variance of the distributions to match the original [3][7]. Another approach to emulate a diverse population is multiple imputation or fractional imputation of single-person observations into different predicted industries and occupations, with person-weights split into a partly synthetic population.

The effort can build iteratively from simple crosswalks, updated with predictions both from analyst studies of particular industries or occupations, and from machine learning methods. Test and benchmark criteria can be introduced over time, tracking quality scores for each augmented dataset.

VI. Conclusion

This would be the first known implementation of a system to jointly impute individual industry and occupation across several Census and CPS data sets based on large scale training microdata and economy-wide benchmarks. 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.

References

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

[2] D. W. Jorgenson, M. S. Ho, and K. J. Stiroh. "Labor Input and the Returns to Education." Chapter 6 of Information Technology and the American Growth Resurgence. Cambridge: MIT Press, 2005.

[3] R. J. A. Little and D. B. Rubin. Statistical Analysis with Missing Data, Second Edition, 2002.

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

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

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

[7] T. K. White, J. P. Reiter, and A. Petrin. Imputation in U.S. Manufacturing Data and Its Implications for Productivity Dispersion. Review of Economics and Statistics vol. 100, pp. 502-509, 2018.