INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
MICHIGAN RETIREMENT RESEARCH CENTER
UNIVERSITY OF MICHIGAN

Promoting research on retirement and Social Security policy

Research Brief 392 | DECEMBER 2018

Feasibility and Reliability of Automated Coding of Occupation in the Health and Retirement Study

Brooke Helppie-McFall and Amanda Sonnega *

Increasing numbers of researchers are interested in using occupational history data from the Health and Retirement Study (HRS) in their research. The detailed data currently are hard to use for longitudinal research, however, because they were coded at different times, and the codes, therefore, are inconsistent over time. The earlier occupation codes provided great detail about manufacturing occupations but little detail about occupations in the service sector. Most notably, they provide no detail on the many new occupations that have appeared as a consequence of computerization and automation. For researchers interested in tracing occupational characteristics or exposures over time, or understanding job ladders and transitions, the current approach generally begins with creation of crosswalks to harmonize codes over time. Due to the differences over time in (1) the existence of certain types of jobs and (2) granularity of occupational categorizations in some areas, large numbers of detailed occupations may end up combined in broad categories that do not allow one to apply different characteristics to different detailed occupations, or to clearly indicate occupational changes over time.

A potential solution to this problem is to recode all occupation and industry data using the current 2010 census coding schemes. All historical and current data would then be compatible and more useful to researchers and policymakers. The availability of consistent occupation and industry codes for all or much of the lifetime careers of HRS respondents would open up many important research questions concerning the determinants and consequences of the changing United States occupational structure. This may be especially useful for studying changes to the cognitive and physical demands of work on the economic security and physical and brain health of workers as they enter retirement.

Hand-coding of occupation and industry, as has been done historically and currently at the HRS, is time-consuming, expensive, and has fairly low inter-rater reliability. This problem is not unique to the HRS and many other researchers have sought potential solutions. With advances in computing, an alternative that has emerged is computer-assisted and/or automated coding. Researchers are increasingly documenting the process of

* ***Brooke Helppie McFall*** *is an assistant research scientist at the Survey Research Center within the Institute for Social Research at the University of Michigan.* ***Amanda Sonnega*** *is a research scientist at the Institute for Social Research at the University of Michigan. This research brief is based on MRRC Working Paper 2018-392.*

computer-assisted and automated occupational coding, generally reporting on comparisons to manually coded occupational data. Interestingly, studies differ on whether they hold the human or the machine as the "gold standard." Some researchers have suggested that the most valid coding may come from a combination of human and machine efforts. It is important to note that intertemporal inconsistency due to crosswalking historical data is only one source of error in occupational data. Other important sources of error include miscoding by coders, which speaks more directly to the initial validity of the coded results.

In this project, we tested the NIOSH Industry and Occupation Computerized Coding System (NIOCCS) to see if it could be useful for coding data from the HRS. NIOCCS is an automated coding engine that translates text to standardized codes. NIOCCS has been continuously updated, and we used NIOCCS v. 3.0 in this project. NIOCCS employs a machine-learning algorithm that uses a large dataset of manually coded occupations as training data for automatic classification. Its stated purpose is to provide a tool that reduces the high cost of manually coding occupation and industry information while improving uniformity of the codes. We then tested the results from NIOCCS against results from a human coder for multiple datasets.

Results of automated reading and coding of HRS data are encouraging. We found that NIOCCS works well only with short descriptions, one to three words each, of job title or job description and "what a business does or makes" as inputs, a finding in accord with recent research. NIOCCS performs reasonably well compared to coding results from a highly-trained, professional occupation and industry coder, with Kappa inter-rater reliability on detailed codes of just less than 70 percent and agreement rates on broader codes of around 80 percent. The main weakness of NIOCCS appears to be its failure to produce codes in many cases. Code rates for NIOCCS for the datasets tested ranged from 60 percent to 72 percent, as compared to a professional coder's ability to code those same datasets that ranged from 95 percent to 100 percent.

NIOCCS may be a useful tool for reducing the human coder hours needed for coding industry and occupation data for the HRS and other studies and datasets. In its current form, it would be most useful as a way to reduce the number of cases a human coder must code, or a way to reduce the amount of time a human coder must spend on each case, or as a first cut for coding historical data that do not crosswalk cleanly to a newer codeframe. For example, with respect to the historical data currently in three different codeframes, we suggest that it would be most cost effective and reliable to do the following:

- use a crosswalk to link only those entries that cleanly map into the newest codeframe for detailed occupations;
- for the remaining entries, use job title and the label for the industry code that was assigned in the historical data to run through NIOCCS; and
- manually code those cases NIOCCS is unable to autocode, then review all autocoded results and crosswalks.

University of Michigan Retirement Research Center
Institute for Social Research 426 Thompson Street Room 3026
Ann Arbor, MI 48104-2321 Phone: (734) 615-0422 Fax: (734) 615-2180
mrrcumich@umich.edu www.mrrc.isr.umich.edu

Sponsor information: The research reported herein was performed pursuant to grant RRC08098401-10 from the U.S. Social Security Administration (SSA) through the Michigan Retirement Research Center (MRRC). The findings and conclusions expressed are solely those of the author(s) and do not represent the views of SSA, any agency of the federal government, or the MRRC.