Comparing Manual and Automated Industry and Occupation Coding: Accuracy and Cost from the Perspective of the California Health Interview Survey

Royce Park[1], David Grant[1], Matt Jans[1]
Marisol Frausto[1], John Rauch[2]
[1]California Health Interview Survey, UCLA Center for Health Policy Research, 10960 Wilshire Blvd, #1550, Los Angeles, California 90024
[2]Westat, 1600 Research Boulevard, Rockville, MD, 20850

Abstract

Assigning industry and occupation (IO) codes to open-ended employment responses can be a time-consuming, expensive, and error-prone endeavor. This study compares manual coding with an automated, computer-assisted coding system developed by the National Institute for Occupational Safety and Health (NIOSH). California Health Interview Survey (CHIS) I&O coding traditionally involves human coders that review and categorize respondent job titles based on verbatim text entries by CATI interviewers. The manual coding scheme uses 2010 Census occupation codes and the 2012 North American Industry Classification System (NAICS). It includes a double-blind process to validate I&O codes followed by adjudication of conflicting codes. The NIOSH Industry and Occupation Computerized Coding System (NIOCCS) uses an automated coding algorithm to assign I&O codes to text entries. A user can submit multiple records (batch-mode) via a Web interface (http://wwwn.cdc.gov/niosh-nioccs/). We randomly selected 1,000 manually-coded cases from 2013-2014 CHIS and processed them with the online NIOCCS system. Preliminary results suggest a clear benefit from using the NIOCCS as it substantially reduces the time and resources necessary to complete the coding, both in person-hours and project duration. Our final analysis compares reliability of each coding system, and assess their success for industry and occupation codes separately. We also discuss cost and data quality trade-offs of each system, as well as operational issues of implementation.

Key Words: industry and occupation coding, automated coding, survey costs, NAICS, SOC

1. Introduction

The California Health Interview Survey[1] (CHIS), like many other general-population surveys, collects industry and occupation data from adult respondents. As a result, CHIS data users can assess health outcomes, health insurance coverage, and other health-related indices as they relate to employment sectors and job types. However, collecting these data is not without complication. Because the commonly-used industry and occupation codes are very detailed, open-ended responses are gathered during the interview, and the data are classified in post-collection processing. This has historically been done by human coders through a process of double-coding and adjudication by a supervisor. With the release of a

[1] http://healthpolicy.ucla.edu/chis/Pages/default.aspx

web-based, "intelligent" NIOCCS system[2], there is new opportunity to reduce the cost and increase the timeliness of this process.

1.1 Industry and Occupation Coding Challenges

"Industry" measures are designed to capture the economic or business sector in which a respondent works. Some examples are health care, agriculture, and construction. "Occupation" measures describe the tasks and activities in which respondents engage at their jobs (e.g. caregiver, administrator, and surgeon). I&O coding schemes are fairly esoteric and complex, so the only reasonable way to capture this data from a respondent in a general-purpose survey is through open-ended responses.

The open-ended responses then need to be categorized into commonly-used I&O coding schemes. There are two population industry and two popular occupation coding schemes that we use in this study. The 2010 U.S. Census Bureau industry categories are based on the 2007 North American Industry Classification System.[3] The 2010 U.S. Census Bureau occupation categories are based on the 2010 BLS Standard Occupational Classification (SOC) system[4].

Coding industry and occupation text is time-consuming, costly, and error-prone. To avoid errors inherent in having a single coder do all the work, and to estimate the reliability of the coding process, many researchers use a double-blind coding process in which two coders independently apply the same coding scheme to a common sample, and discrepancies are adjudicated by a third coder or supervisor. Clearly, this adds significant time and resource expenditure to the coding process, and can lead to usable data being released long after it was collected.

Automated coding, like that developed by NIOSH has several benefits. The first, and most obvious benefit is speed. Second, the automated coding process uses all information in its database, which comes from initial training by expert coders, and users who have processed their data through the system. It is essentially an intelligent system, which learns and refines its coding accuracy the more it is used. Finally, the system produces an estimated reliability, something that could not be obtained from a single coding pass by a single coder, but which is obtained through processing the data in NIOCCS one time. Interested readers should consult the NIOCCS documentation[5] for more information about its reliability calculation.

Figure 1 delineates the NIOCCS coding engine. As a batch of data is input, it is run through a tiered coding process including a word proximity and phonetics match. Confidence level preferences and restrictions are then utilized to produce auto coded records.

[2] http://wwwn.cdc.gov/niosh-nioccs/
[3] http://www.census.gov/eos/www/naics/
[4] http://www.bls.gov/soc/
[5] http://www.cdc.gov/niosh/topics/coding/overview.html

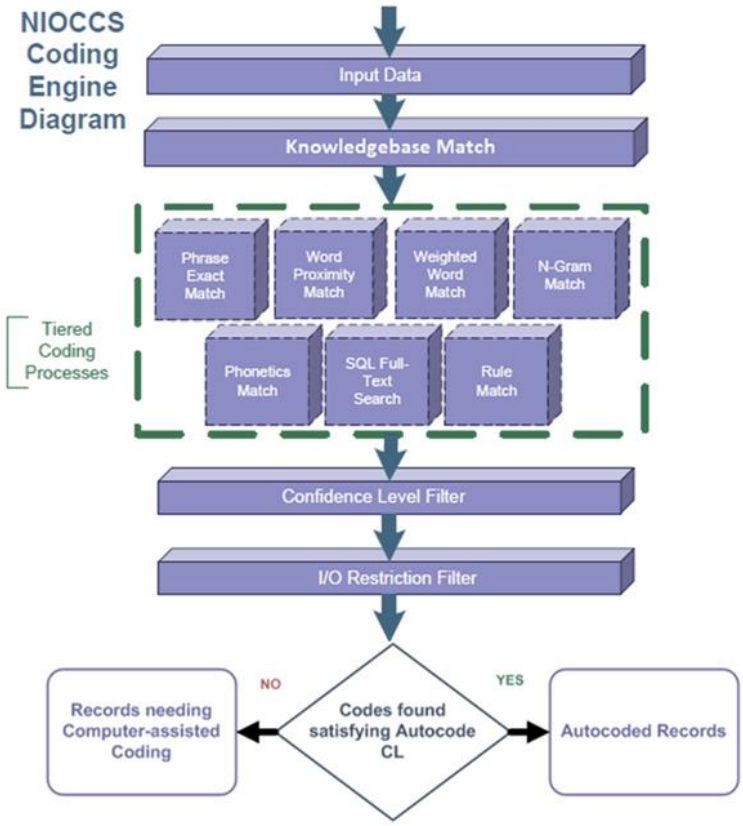

Figure 1. NIOCCS Coding Engine
Source: National Institute for Occupational Safety and Health Division of Surveillance, Hazard Evaluations, and Field Studies, 2015

1.2 Research and Application Question

This paper assesses the concordance (i.e., inter-rater reliability) of industry and occupation (I&O) data coded by human and automated systems using two sets of commonly-used I&O codes. We asked whether an automated coding system could a) produce results faster than human coding**,** b) produce similar results to human coding, and c) reduce costs.

2. Methodology

2.1 Data Source

A random sample of 1,000 cases reporting I&O data in CHIS interviews conducted by Westat between January 2013 and January 2015 as part of the 2013-2014 CHIS data collection cycle (of 18,739 employed respondents) was used for this study. Respondents were asked the following questions about their job during the phone interview, and verbatim responses were recorded in the CHIS computer-assisted telephone interviewing (CATI) system.

A) "On your main job, are you employed by a private company, the government, or are you self-employed, or are you working without pay in a family business or farm?" (Question AK4)

B.1) "What kind of agency or department is this?" [PROBE FOR AND RECORD BOTH THE LEVEL OF GOVERNMENT (E>G., STATE, LOCAL) AND THE FUNCTION (E.G., BUDGET OFFICE, POLICE, ETC.] (Question AK5)

-OR-

B.2) "What kind of business or industry is this?" [IF NEEDED, SAY: "What do they make or do at this business?'] (Question AK5)

C) What is the main kind of work you do? (Question AK6)

Questions B.1/B.2 (AK5) provided the industry responses and C (AK6) provided the occupation responses. B.1 was used if the respondent worked in government, and B.2 was used if the respondent worked for a private company or was self-employed. If the respondent had more than one job, their "main job" was defined as the one at which they worked the must hours. See the entire CHIS questionnaire is online.[6]

2.2 Coding Methods and Coding Schemes

2.2.1 National Institute for Occupational Safety and Health: Industry & Occupation Computerized Coding System (NIOCCS)

The NIOSH Industry and Occupation Computerized Coding System (NIOCCS) is a web-based software tool designed to translate industry and occupation (I&O) responses into standardized I&O codes. Its primary benefit is reducing the high cost of manually coding I&O data, and has the additional benefit of improving uniformity of the codes, relative to human-coded data. System features include: a) completely automatic coding in which the user can enter a single text responses or upload batch files of multiple cases of raw data and receive coded responses back; b) defined levels of code reliability, which are based on all the cases previously submitted to the NIOCCS system; c) "computer-assisted" coding in which the system offers likely code options for cases that it cannot code at a specified level of reliability; and d) coded data provided various coding schemes (e.g., "crosswalked coding"). NIOCCS is available free of charge (users should register for a free account) and requires only internet access and a web browser for use.

[6] http://healthpolicy.ucla.edu/chis/design/Pages/questionnairesEnglish.aspx

2.2.2 NAICS and SOC coding schemes

The U.S. Census Bureau currently collects data on industry, occupation, and class of worker for Americans in the labor force on several surveys. For this study we utilize the 2012 North American Industry Classification System (NAICS) codes and the 2010 Census occupation codes.

The North American Industry Classification System (NAICS) is a system for classifying individual business locations, or establishments, by type of economic activity in the United States, Canada, and Mexico. The system consists of 270 categories arranged into 20 sectors. Its purposes include: 1) facilitating the collection, tabulation, analysis and presentation of data; 2) promoting uniformity and comparability in the presentation and analysis of statistical data that describes the North American economy. NAICS is widely used by Federal statistical agencies, state and local agencies, trade associations, private businesses, and other organizations. NAICS is a 2- through 6-tiered classification system, offering five levels of detail. Each code digit corresponds to a continuum of progressively narrow categories. More digits demonstrate greater detail in classification. The first two digits represent the economic sector, while the third digit designates the subsector. The fourth displays the industry group, the fifth displays the specific NAICS industry, and the last digit designates the national industry. 5-digit codes represent the level that allows comparability for NAICS sectors across the three countries, while 6-digit codes represent complete and valid NAICS codes.

The 2010 Census Occupation classification codes is a system for classifying individual's main or secondary employments, developed by the Standard Occupational Classification (SOC) Manual. The occupations in the SOC are designated by a six-digit code, classified at four levels of combination: major group, minor group, broad occupation, and detailed occupation. Each lower level of detail identifies a more specific group of occupations. The first two digits represent the major group, while the third represents the minor group. The fourth and fifth represent the broad occupation, and the last digit represents the detailed occupation. Occupation codes always end with the digits 0 through 6. SOC has 509 separate categories arranged into the 23 major groups. There are 23 major groups that are divided into 97 minor groups, 461 broad occupations, and 840 detailed occupations.

2.2.3 Comparisons

The primary comparison in this project was between human-coded cases (conducted by Westat), and cases processed through the NIOCCS online system. A fraction of the NIOCCS-processed codes had to also be human adjudicated by selecting from among the recommended codes the NIOCCS system provided when it could not automatically code them (note: this is called "computer-assisted" in the NIOCCS system to distinguish it from auto-coded results, but it essentially involves human intervention and judgement.

2.2.3.1 Human-coded "gold-standard"

The historical method of coding CHIS I&O data was to have a human coder code the open-ended text, with discrepancies or confusing cases adjudicated by a second trained I&O coder. The 1,000 sampled cases were first coded through this method and serve as our comparison. The manual coding scheme includes a double-blind process to validate I&O codes followed by adjudication of conflicting codes.

Using the NIOSH/NIOCCS online single record coding scheme, sets of given I&O responses are entered. Often times, the output contains multiple coding options, spanning

low to high reliability threshold. In clear cases, the coding is selected at the highest, 90% or higher, confidence level threshold. In disputable cases, there is adjudication of conflicting codes. This process ensures 100% double coding through the process of settling questionable codes.

2.2.3.2 Automated coding

The automated coding was facilitated through the online NIOCCS (NIOSH Industry & Occupation Computerized Coding System). The software utilizes the Census 2010 Classification scheme for I&O to produce codes at selected levels of reliability (e.g., high [≥90%], medium [≥ 70%]). We used the high reliability setting for our auto-coded data.

While the majority of data input can be auto coded, depending on the confidence level preferences, extraneous and unidentifiable cases are revealed. These non-auto-coded responses can be coded manually or with computer-assistance based on suggestions from NIOCCS system. 50% are non-auto coded at "high" (Residual), while 30% are non-auto coded at "medium" (Residual). The "residuals" were manually coded with computer-assistance by two of the authors.

The NIOCCS automated coding process, including partial computer assisted coding, was completed in 30 hours. Selected at a high coding threshold, the NIOCCS system coded a large portion of the random sample, while the remaining cases were human-coded though computer assistance. A complete coding through NIOSH with half computer-assisted coding takes 30 hours; completed coding for the same sample, through two computer-assisted coders, would yield around 60 hours.

This NIOSH/CDC product is available for free online.[7]

2.1 Reliability Assessment

The reliability assessment compared the I&O manual and automated coding outputs. The 1,000 randomly-sampled cases came from the CHIS 2013-2014 interviews. The agreement rate, or percent of cases with agreement between human and NIOCCS coding schemes, determines the reliability of manual coding. The NIOCCS auto-coding results are based on the high-reliability results mentioned above. NIOCCS auto-coded residuals (i.e., non-auto coded) are human coded in the NIOCCS system, assisted by NIOCCS suggestions.

3. Results

Figure 2 shows the simple agreement (i.e., concordance) rates for the industry codes (left-most 8 bars) and occupation codes (right-most 8 bars). The rates are further broken down by whether we compared the first four digits ("4-digit (detailed)") or the first two digits ("2-digit (major)") of the coded data. The two digit (major) code represents a broad classification of a specific industry (i.e. Education). The four digit (detailed) code represents a detailed, narrow subsection of a larger industry selection (i.e. Grade School Education).

For each type we also compared human coding by Westat with auto-coding by NIOCCS (i.e., only those that were completely auto-coded), human coding by Westat with the

[7] http://wwwn.cdc.gov/niosh-nioccs/

human-coded "residual" (i.e., "computer-assisted" cases using the NIOCCS recommendations but with human decision). We also present agreement of the overall codes (i.e., Westat's human-coded with all the NIOCCS-coded combined).

As expected, two-digit summaries were more reliable simply because the coded data were split into fewer categories. In both the four-digit and two-digit summaries, the highest agreement rate for both industry and occupation was for the human coding v. auto-coding (56% and 62% for four-digit industry and occupation respectively, and 87% for both industry and occupation when summarized to the two-digit level). Agreement was lowest for the Westat human-coded using the traditional method v. computer-assisted human-coded residual using the NIOCCS system's suggestions.

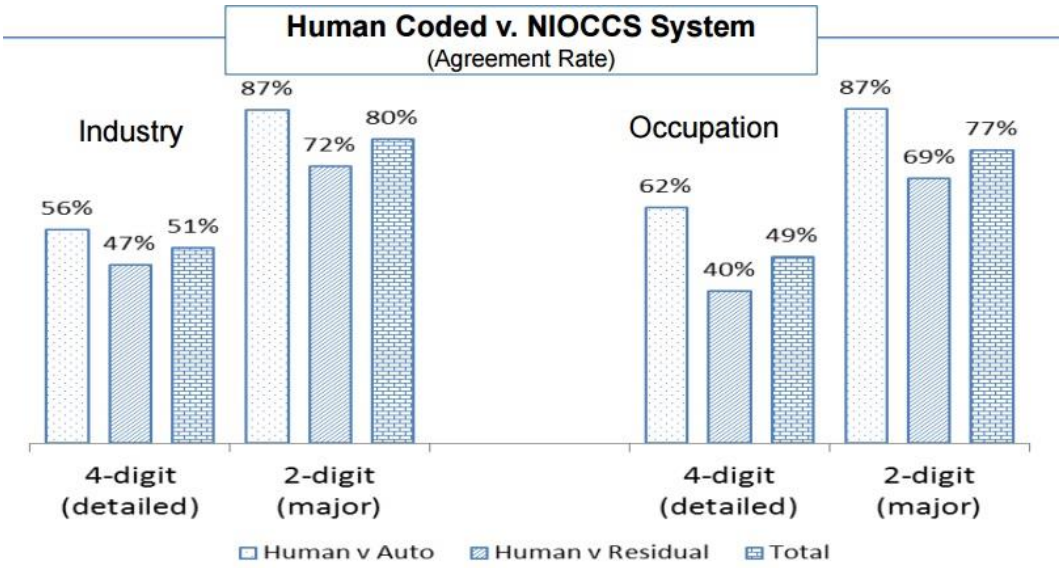

Figure 2. Manual v. Automated Coding Agreement Rate

This comparison reinforces auto-coding as a great source to reduce costs, efforts, and overall duration of the coding process. By facilitating the reduction of time completion and human time efforts, project costs are reduced. It is evident that automated coding is beneficial for small or quick-turn-around projects.

While automated coding can substantially reduce costs, efforts, and project duration, it does not ensure 100% complete and accurate coding. Many cases still require manual intervention. Discordant cases reveal the subjective nature of coding.

4. Discussion

Given the complexity of the data and coding system, the overall agreement rates seem reasonable, at least at the two-digit level. It is not surprising that agreement goes down as we split the cases into more detailed codes.

There are several important caveats to this analysis. The largest is probably the nature of truth. Nearly all reliability assessments will have agreement less than 1.0. As in most

reliability assessments, which measure is "truth" or the "gold standard" relies more on argument and purpose than on anything objective. Because we had traditionally used full human coding, that was a reasonable "gold standard" for evaluating a new approach. However, that does not guarantee that the human-coded results are more accurate than the NIOCCS-coded results. An argument can be made that the NIOCCS system, with its massive database of 1.26 million results informing each classification, and having been developed by experts in I&O coding, is the true gold standard. Obviously, the true values of these codes are not known in this case. However, it could be tempting to assert that the automated coding is less reliable (i.e., only in 77-80% agreement with the human coding), but that would be assigning truth to the human-coding method simple because of its historical precedent. Clearly, a third data source would be helpful in assessing the accuracy (not just the reliability) of the codes.

There are some other limitations to a complete assessment of the reliability of these codes. First, Westat's human coding was not assessed for reliability within itself. Second, we present agreement rates, not Cohen's Kappa or similar reliability statistics that control for chance agreement. In analyses not reported here, the relative ordering of Kappa is about the same as the ordering of agreement rates in Figure 2, although obviously lower. We also considered weighted Kappa options that put less weight on small discrepancies in the detailed code digits (e.g., "computer technician" v. "IT technician") and more weight on discrepancies in the first few digits (e.g., "computer technician" v. "teacher"). The true reliability profile of these data is complex.

As mentioned previously, the NIOCCS automated coding process, including partial computer assisted coding, was completed in 30 hours. Fast-paced automated coding in combination with computer- assisted coding reduces the time needed to complete extensive coding in comparison to full human-coding. It would take about double the time, or 60 hours for two individuals to manually code the random sample of 1,000 cases. Two coders would be necessary to ensure accuracy and comparison measures. A reduction in time (human-hours and project duration) and resources is a clear benefit from using the online system.

Acknowledgements

We thank the CHIS coding staff at Westat for their hard work on this project. We also thank the CHIS sponsors in general, who can be found online at: http://healthpolicy.ucla.edu/chis/about/Pages/funds.aspx