WDQI Indiana

Knowledge Transfer

Indirect Method
of Occupational Assignment

Indiana Department of Workforce Development &
Indiana Business Research Center at the
Indiana University Kelley School of Business

Introduction

The goal of this occupation assignment module is to take the unassigned cases of wage records and assign SOC codes based on the statistics of assigned cases. This is what we call the imputation method. There are two stages in this assignment. In stage one, we attach SOC codes from administrative records (direct assignment) to the wage records. This stage prepares data for the second stage, where a *random assignment* is performed. It is called random assignment, since it assigns SOC codes to randomly-selected records based on statistic information—involving random probability, opposed to the exact match based on certain criteria.

The assigned cases come from four sources, Occupational Employment Statistics (OES) micro administrative data, public employees, public licensing agency, and UI claims. Expect for UI claims that self-report SOC codes, SOC codes from the other agencies were imputed separately. Thus, it is possible that a record cross-listed with different agencies bears different SOC codes. For example, a licensed nurse (public licensing) who applied for unemployment insurance (UI claims) in 2103 had worked in a state hospital (public employee). Based on the "error" associated with each data source, we rank public licensing as the most trustworthy source, followed by public employees, OES micro data and UI claims. Thus, when disagreement on a SOC code occurs, we went by with the source in this rank order.

Wage Records Overview

The administrative data we received from DWD has 3,385,473 unique records with valid NAICS codes in 2013. It reports the top three jobs—based on their salaries—that an employee had worked in that year. 68 percent of employees have only one job, 27 percent have two jobs and 5 percent have three. Since majority of employees had worked for one job and the dominant job, which is defined as the job with the highest earnings share of the total wages received in that year, on average accounts for 92 percent of total salaries received, we focus on the dominant job. After removing the "minor" jobs, there are about 1 percent of duplicated records—employees had worked on multiple jobs (different NAICS codes) with equally "highest" pay. We kept those duplicates and treated them as multiple records. Thus, a record is uniquely defined by a record's ID (i.e. universal ID) and NAICS code.

The wage records also contain demographic information, such as date of birth, gender, race and educational attainment. However, this information is not universally available: 80 percent of records were recorded with date of birth, 55 percent with gender, 54 percent with race and only 10 percent report education. Although demographic information may facilitate the SOC assignment, we did not use this information due to missing data. Thus, the random assignment takes only industry (NAICS), wages and location as the input from the wage records.

Merge data sets

In stage one, we need to attach the SOC codes from various sources to the wage records. Here, we go through each of the four data sources, in terms of the characteristics of their SOC codes and how they were mapped to the wage records.

Public Licensing Assignment (PLA)

assigns 6-digit SOC codes to individual records holding public licenses. This assignment is based on a license-SOC crosswalk, which matches a license with its all possible SOC codes. (for details on this assignment, refer to Public Licensing Agency Data). Since PLA is a probable matching system, a record may be assigned with multiple SOC codes. In deciding which SOC code to keep, we followed a rank scheme for reliability: higher ranks—thus more reliable assignment—are given to more specific matches (e.g. 6-digit codes having higher ranks than 5/4/3-digit) and to more precise matches (e.g. "nurse practitioner tiebreak" having higher ranks than "25-75% wage match"). We retained the most "reliable" SOC codes in the ascending order of ranks (see **Table 1**). In the final PLA data, 67 percent of SOC codes were matched via "6-digit direct" (the 1st rank), and nearly 90 percent of SOC codes were matched through the top two ranked methods.

Table 1: Rank of assignment method for PLA

Rank	Assignment method
1	6-digit direct
2	6-digit alcohol wage tiebreak; 6-digit nurse practitioner tiebreak; 6-digit asbestos tiebreak; 6-digit engineering wage tiebreak; 6-digit partial quarter tiebreak
3	5-digit direct
4	5-digit partial quarter tiebreak
5	4-digit direct
6	4-digit partial quarter tiebreak
7	3-digit direct
8	3-digit engineering wage tiebreak; 3-digit partial quarter tiebreak

9	6-digit 25-75% wage match;
	5-digit 25-75% wage match;
	4-digit 25-75% wage match;
	3-digit 25-75% wage match;

10	6-digit 10-90% wage match;
	5-digit 10-90% wage match;
	4-digit 10-90% wage match;
	3-digit 10-90% wage match;

The top industries holding public licenses are healthcare and social assistance (51 percent), accommodation and food services (18 percent) and administrative support and civil services (7 percent); the top occupational groups holding public licenses are healthcare practitioners and technicians (42 percent), food preparation and serving (23 percent) and healthcare support (21 percent).

Public Employees (PE) assigns 6-digit SOC codes to individual records based on job titles and the departments they worked in (information on industrial NAICS codes and location are unavailable for this data set). Since the assignment is entirely based on text description, records that worked in various departments and under various job titles could be assigned with multiple SOC codes. (For details on this assignment, Public Employee Data.) In such cases, we kept the highest paid SOC code based on a record's total compensation from each job. This gesture is consistent with the treatment on wage records (see the previous discussion). In this final data set, 94 percent of records were directly matched on job titles and the rest were indirectly matched on department or a combination of job title and department.

When merging the PE data set with wage records, we used only ID as the merging criterion, opposed to a combination of ID, NAICS and county fip codes used in other data sets. The problem with this matching is that the assignment of SOC to NAICS codes are not necessarily one-to-one. In such cases, manually sorting them out is quite burdensome. Thus, to select the "best" NAICS code that a SOC code is assigned to, we calculated the (absolute) deviation between annualized NAICS wages (from wage records) and SOC compensation (from PE) and assigned the SOC code to the NAICS, with which the income gap is the smallest.

The 2013 PE-matched wage records show that education (63 percent), government (22 percent) and healthcare (5 percent) are the top three industrial sectors that employ public employees. Accordingly, the top three occupational groups for public employees are education (40 percent), administrative support (16 percent)) and protective service (11 percent).

OES Microdata assigns 6-digit SOC codes to individual records based on OES micro staffing patterns. A wage record with a match to the microdata, based on employer's UI account ID and wage information, has a list of narrowed-down possible SOC codes for those quarters he/she worked for that employer. The procedure also incorporates the PLA data when possible to further narrow down the list of occupations. (For details on this assignment, refer to OES Microdata.) An employee might be assigned with multiple SOC codes for the same industry he/she had worked in, because the employee had either switched employers from time to time or worked for various employers at the same time. The latter case happens when, for example, in healthcare a physician is in affiliation with various healthcare facilities (e.g. hospitals); a registered nurse works in various physician's offices. To ensure the match between NAICS and SOC codes are one-to-one, we kept the SOC codes with the highest total compensation.

The 2013 OES-matched wage records in the healthcare industry show that healthcare practitioners and technicians (35 percent), healthcare support (16 percent) and office and administrative support (15 percent) are the top three occupational groups.

UI Claims record, for each applicant, a voucher entry date, benefit end date, industrial 6-digit NAICS codes, occupational 2-digit SOC codes and employer's county fip codes that one had worked for the previous job, as well as basic demographics (such as birthdate, gender, race, education and zip codes). We used voucher entry date to approximate working quarters, assuming an applicant filed for UI immediately after he/she became unemployed.

We were able to link UI claimants with their wage records via a universal ID. A record was assigned with the 2-digit SOC code from UI claims as long as its NAICS code, county fip code and working quarter match. For 2013, there are 173,789 (NAICS-fip-quarter) records were assigned with unique 2-digit SOC codes from UI claims. The industries that filed the most UI claims are construction (24 percent), production material manufacturing (17 percent) and administrative support and civil services (12 percent). Accordingly, the occupations that filed the most UI claims are construction (18 percent), production (17 percent) and transportation and material moving (16 percent).

We realize that each data source may bias towards certain occupations. For example, PLA largely represents the healthcare sector, in which certification is standard; PE largely represents workers in educational services, in which public schools/colleges play a big role in the state; UI claims largely represents low-income and low-skilled workers, since they are more prone to lose jobs and apply for UI benefits. Thus, relying on information of these SOC codes would also introduce bias to our results. After the first stage, the overall SOC "recovery" rate for the 2013 wage records is about 20 percent. On the other hand, if taking the healthcare industry alone, the recovery rate is 75 percent. Vast majority of SOC codes for healthcare workers come from OES micro data (52 percent) and PLA (44 percent).

Random Assignment

In this stage, we assigned SOC codes to remaining records based on state and national industry-occupation distributions. This procedure starts with assigning broad (2-digit) SOC groups and then works down to 6-digit codes under each of the 2-digit SOC groups. This assignment was done separately for each county and involves the following three steps.

Step 1: Calculate population controls for the number of possible SOC codes within an industry.

We first calculated the occupational staffing pattern—shares of occupations within an industry—based on the state OES micro data and the National Industry-Specific (available from BLS) profile for sectors uncovered by the micro data. For each industry, we then ranked occupations—from the most likely to the least—according to their staffing pattern.

Next, using IBRC's "no-holes" QCEW estimates (i.e. industrial headcount) and the occupational staffing pattern we calculated the total headcount (by *QCEW* × *staffing pattern*) of SOC codes within each industry. Note that the staffing patterns for 2-digit SOC codes are based on 4-digit NAICS codes, and for 6-digit SOC codes based on 5-digit NAICS codes.

Finally, we calculated total SOC headcount minus headcount from assigned cases to get the SOC "allowance", used as population control, for remaining wage records.

Step 2: Calculate the wage distribution of SOC codes within each industry.

The wage distribution of SOC codes was approximated by the log-Normal distribution. We calculated the statistics (i.e. average and standard deviation) of wages from already assigned SOC codes, and used them to approximate the wage distribution of potential SOC codes for unassigned records. Note that assigned SOC codes sometimes can be too sparse to produce statistics. In such cases, we substituted with the wage statistics from the National Industry-Specific profile.

We expanded remaining wage records with all potential SOC codes, along with their wage statistics, for the industry they worked in. Next, for each record we calculated the probability density of its wages under each of the potential SOC codes—that is, how likely a record worked in particular SOC industry given the wages it earned. Larger density values mean more likelihood. Thus, we were able to rank the potential SOC codes from the most "favorable" (most likely) to the least, based on their probability densities. We only retained the top fifteen SOC codes for each record.[1]

Step 3: Random assignment

For each industry starting with the most likely SOC code (e.g. healthcare practitioners and technicians in the healthcare industry), we randomly drew records from the pool of candidates, whose most favorable SOC codes match with that of the industry's, until the SOC headcount reaches its population control.

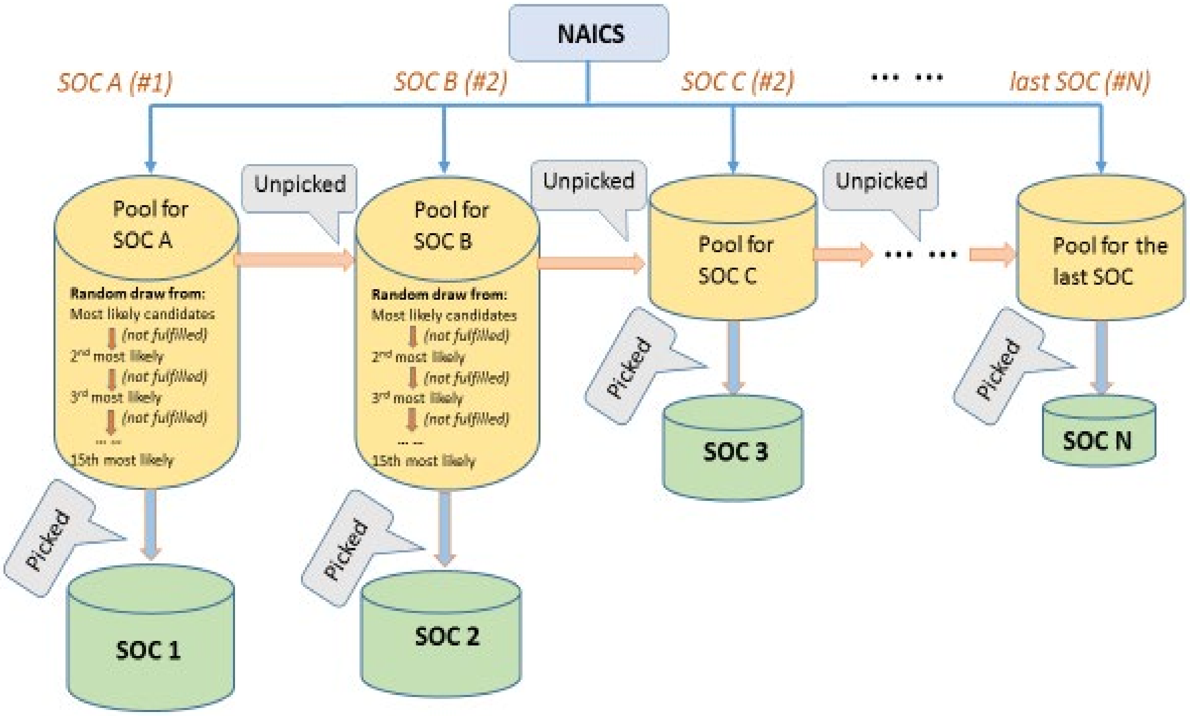
When a SOC's population control exceeds the headcount of candidates, essentially everyone in the group was assigned with that SOC code, and these records were removed from future draws. When a SOC's population control does not meet the headcount of candidates, there will be "leftovers" and remaining candidates will enter future SOC draws in their less-favorable groups, that is, the 2nd, 3rd, 4th, …, 15th most likely SOC group. **Figure 1** shows the process.

Use the healthcare sector—hospitals—as an example. The most common occupational groups in hospitals are healthcare practitioners and technicians (e.g. physicians and registered/licensed nurses), administrative support (e.g. office clerks and receptionists), healthcare support (e.g. health aides), … Based on their salaries, employees who work in hospitals are in favor of different occupations. For example, employees who earn 300k a year are in favor of physicians, and those who earn 30k a year in favor of office clerks. We start with registered nurses, the most common occupation in hospitals. We randomly draw individuals whose favorable occupations are registered nurses and assign them the SOC

[1] The number fifteen is arbitrary. However, as the values of probability density tend to be very small for majority of SOC candidates—signaling very unlikely, we think it's ok to ignore those ranked at the bottom.

code. These individuals are then removed from future draws. For those who did not get picked will enter the next run of their second favorable occupation, say therapists. If they get picked on the second run, end of story; if not, they will enter the 3rd, 4th, 5th, …, until the 15th run. In each run, we always start with the most common occupation in the hospital sector (i.e. registered nurse) and cycle down to the least common occupation (e.g. nurse midwives). Only individuals whose *current* favorable occupation match with that of the hospital's will enter the draw.

Figure 1: Random assignment workflow



Results

We used 2013 wage records as our test sample. The remaining wage records contain 1,120,073 unique industry-county specific cases, about 33 percent of total records received. The recovery rate is roughly 40 percent. This low rate is because many SOC codes have no allowances (i.e. zero population control) to start with and thus never got picked in the process. If restricting to SOC codes with positive headcounts, the recovery rate is roughly 75 percent. At industry level, over 90 percent of assigned SOC codes came from an industry's top three occupations (out of total 22 occupational groups), and of them 80 percent were from the top one list. At record level, 64 percent of assigned SOC codes came from a record's top five "favorable" occupations (recall each record has a list of 15 favorable SOC codes to select from). However, the most favorable occupations did not always get picked, but the 2nd-4th favorable ones were the most commonly selected.

For example, for the manufacturing sector the recovery rate (with positive headcount) is roughly 80 percent, out of 313,641 unique records. 92 percent of SOC codes came from the industry's top three occupations (out of total 19), and 60 percent from individual's top five occupations. For administrative and support and related services, the recovery rate is roughly 66 percent, out of 172,117 unique records. 94 percent of SOC codes came from the industry's top three occupations (out of total 22), and 74 percent from individual's top five occupations. For the healthcare sector, the recovery rate is roughly 66 percent, out of 43,328 unique records.[2] 74 percent of SOC codes came from the industry's top three occupations (out of total 21), and 73 percent from individual's top five occupations. In healthcare, the top three major occupational groups among remaining wage records are office and administrative support (24%), personal care and community (18%) and social services (10%).

[2] Notice that the total cases for the healthcare sector is small. This is because many occupations in healthcare require professional licenses and many employees work in state hospitals that fall in the category of public employees. Thus, for majority of records their SOC codes were assigned *manually* via other direct methods. For details, refer to the two modules Public Employee Data and Public Licensing Agency Data under Knowledge Transfer at Hoosiers by the Numbers (http://www.hoosierdata.in.gov/wdqi/knowledge-transfer.asp).