Knowledge Transfer WDQI

Occupational Employment Statistics Microdata

Indiana Department of Workforce Development &

Indiana Business Research Center at the

Indiana University Kelley School of Business

HOOSIERS BY THE NUMBERS

Your premier source for labor market information for Indiana.

IBRC

Authors: Ping (Claire) Zheng, SunJung Yoon, Thea Evans, Kellie McGiverin-Bohan

Contributors: Carol Rogers

Acknowledgments

This manual was prepared as a part of an ongoing collaboration with the Indiana Department of Workforce Development and Indiana Business Research Center to provide insights on how to best approach using occupational employment statistics data. We are thankful to have this opportunity from the U.S. Employment and Training Administration as part of the Workforce Data Quality Initiative.

Suggested Citation:

Zheng, P., Yoon, S., et al. Indiana Business Research Center (2019). *Knowledge Transfer WDQI: Occupational Employment Statistics Microdata*. Bloomington, IN: Hoosiers by the Numbers, Indiana Business Research Center, the Indiana Department of Workforce Development.

Contents

Figures

Tables

Contact Information

For more information about this report, contact the Indiana Business Research Center at (812) 855-5507 or email ibrc@iu.edu.

Background

Occupational Employment Statistics (OES) microdata assignment project is designed to use employer wage data to identify an individual's Standard Occupational Classification (SOC) codes in another data set. To accomplish proper wage matches across the two data sets, we categorize employees' wage records in a given year and examine the distribution of the SOC codes and employers. We use various techniques for identification, and there are many steps and additional sources to accomplish matching. In the following sections, we describe the general background of the occupation system, OES, and microdata management system. Then, we explain the details of the knowledge transfer process, from data cleaning to data procedure.

The American Industry and Occupation Classification Systems

The U.S. federal government uses two types of standards to classify industries and occupations. One is the North American Industry Classification System (NAICS) and the other is the SOC codes. Over time, the system is updated to match the current structure of the economy. One difference between the two systems is that the SOC is an internal U.S. federal system grouped by the nature of the work performance, whereas the NAICS is a general classification system across three countries, the United States, Canada, and Mexico. Both systems are used to collect and analyze statistical data for the business economy.

Occupational Employment Statistics Survey Background

The semi-annual OES survey program is conducted by the U.S. Bureau of Labor Statistics (BLS), for nonfarm business establishments to estimate employment and wages by occupation at the 3-digit, most 4-digit, and selected 5- and 6-digit NAICS sector levels. The program produces the estimates for about 800 occupations (excluding self-employment), at the national, state, metropolitan and nonmetropolitan area levels, as well as by industry and ownership. In total, the BLS collects a sample of 1.2 million establishments over the course of three years. The OES data is widely used for different analyses such as occupational employment, wages, occupation developments projection, vocational counseling, skill studies, and market analysis. The survey has traditionally used NAICS to estimate employment and wages; however, the estimates began using the 2010 SOC system beginning in May 2010. For more information on the OES, visit www.bls.gov/oes/oes_emp.htm.

The OES Microdata Assignment System

The OES microdata assignment system adds an additional scope to the SOC assignment as it provides more information in the private sector. The linkage between OES microdata and wage records data can be established via employer's Unemployment Insurance (UI) account number[1] and 6-digit NAICS code. Although an employee's exact SOC codes are unknown, OES microdata has the total number of employees for an occupation. With the given information, assigning potential SOC codes to employees is feasible (see Figure 1).

[1] The Unemployment Insurance programs benefit unemployed workers who meet eligibility requirements.

For some business entities, the type of occupation and industry is straightforward—for example, cab drivers for a cab company, and cashiers for a liquor store—and there are usually only one or two SOC codes in a single industry. Once you link the two data sets by the employer UI account number and 6-digit NAICS code, assigning an employer's SOC codes from the OES data to employees in wage records data is not the main challenge. The main challenge is dealing with large business entities. For instance, a pharmaceutical company, such as Eli Lilly and Company, has two types of businesses (or industries): pharmaceutical manufacturing and research and development. Thus, there are 16 occupations within pharmaceutical manufacturing and 106 occupations within research and development. The goal is not to assign correct SOC codes but rather best assign these SOC codes to the highest probability codes for employees by inspecting wage records data.

To facilitate the mapping, we used the national industry-occupation staffing pattern published by the BLS to classify from the most to the least common occupations by wages within each industry. This provides a ranked profile of all possible SOC codes and serves as a guideline for the SOC assignment. After linking the employer UI account number and 6-digit NAICS codes in the two data sets, we went down the list from the most common to the least common occupations to assign SOC codes to employees based on their wage categories.

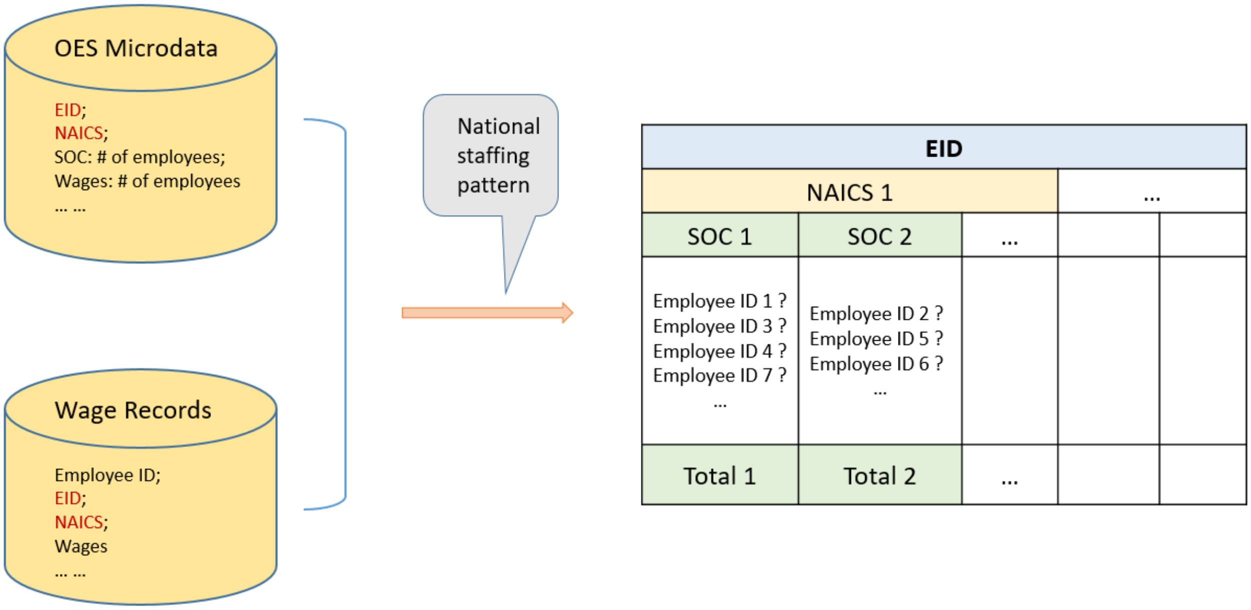


Figure 1: OES Microdata Assignment Overview

Data Overview

There are four key main data inputs in this assignment (Year 2013):

1) OES matched wage records data
2) OES microdata
3) OES national staffing patterns
4) License-SOC crosswalk

Supporting data:

5) Licensing IWIS to IBRC
6) Location Counts – OES and EQUI

The OES matched wage records data and OES microdata are administrative data from the Indiana Workforce Intelligence System (IWIS) generated by the Department of Workforce Development (DWD). The subset of employee wage records are identified in the OES microdata by employer's UI account number. We refer to them as the OES employees.

The OES national industry-occupation staffing patterns data is from the BLS website (look under Subjects, Employment by Occupation, section OES Data and item National). The staffing patterns report all possible occupations, their estimated employment counts and annual wages associated with each industry. The industrial categorization are based on 4-digit NAICS codes.

The License-SOC crosswalk is an Indiana Business Research Center (IBRC) in-house research tool, a bridge file between license and occupation. The mapping between licensure and SOC codes is multiple-to-multiple. For example, there are 30 different licenses associated with Wholesale and Retail Buyers (SOC 13-1022); license title Professional Engineer is associated with 15 different SOC codes. In addition, we incorporated the Public Licensing data, which were used to narrow down the list of occupations when an occupation requires a license. Tables 1-5 describe the key data variables.

Table 1: Employee Wage Records Field Description

Field name	Description
universal_id	Unique identifier for each SSN in the wage records
cnty	State county FIPS codes, numeric codes defined in U.S. Federal Information Processing Standard Publication ("FIPS PUB") 5-2 to identify U.S. counties and certain other associated areas.
uiacct_id	Employer's UI account ID, which has been de-identified by the Department of Workforce Development.
wyear	Year for which wages are being reported
wquarter	Quarter for which wages are being reported
wageamount	Amount of wages paid to employee during quarter

Table 2: OES Microdata Field Description

Field name	Description
uiacct_id	Employer's UI account ID, which has been de-identified by the Department of Workforce Development.
bmkwgt	The representative number (ratio) of the employment for sampled location and how the industry for that location is represented (imputed) in an area.
estimpflag	Anything nonzero is an imputed location.
soc2010	The Standard Occupational Classification code.
sumemp	The number of total people in that particular SOC code for that location.
naics6	The 6-digit NAICS code.
cntyfips	Indiana county FIPs codes. Code '999' is multiple-county residence.

Table 3: OES Wage Categories

Wage category	Salary range
A	Under $19,240
B	$19,240 - $24,439
C	$24,440 - $30,679
D	$30,680 - $38,999
E	$39,000 - $49,919
F	$49,920 - $62,919
G	$62,920 - $80,079
H	$80,080 - $100,919
I	$100,920 - $128,959
J	$128,960 - $163,799
K	$163,800 - $207,999
L	$208,000 and above

Table 4: OES National Staffing Pattern

Field Name	Description
naics	North American Industry Classification System code for the given industry
naics_title	North American Industry Classification System title for the given industry
occ_code	The 7-digit SOC code, or OES specific code for the occupation
occ_title	SOC title or OES specific title for the occupation
occ_group	The SOC occupation level. total=Total of all occupations; major=SOC major group; minor=SOC minor group; broad=Broad SOC occupation; detailed=Detailed SOC occupation level
tot_emp	Estimated total employment rounded to the nearest 10 (excludes self-employed)
emp_prse	% relative standard error for the employment. Relative Standard Error (RSE) is a measure of the reliability of a statistic; the smaller the relative standard error, the more precise the estimate.
pct_total	% of industry employment in the given occupation. Percentage may not total to 100 due to occupational data not reported.
pct_rpt	Percent of establishments reporting the given occupation in the given industry
h_mean	Mean hourly wage
a_mean	Mean annual wage
mean_prse	% relative SE for the mean wage. Relative Standard Error (RSE) is a measure of the reliability of a statistic; the smaller the relative standard error, the more precise the estimate.
a_pct10	Annual 10th percentile wage
a_pct25	Annual 25th percentile wage
a_median	Annual median wage (or the 50th percentile)
a_pct75	Annual 75th percentile wage
a_pct90	Annual 90th percentile wage

Table 5: IBRC Bridge File

Field Name	Description
ONETTITLE	The Occupational Information Network (O*NET) occupational definitions.
soc2010	2010 definition of Standard Occupational Classification system.
title2010	2010 definition of SOC code titles for occupations.
board	General title derived from profession names.

Figure 2 presents the relationship among the various data components. The red boxes are the four main data inputs, the green box is Public Licensing data—a parallel work under SOC assignment engine—and the blue boxes are merged data sets. What we are essentially doing is to first, build an occupation profile consisting of all possible SOC codes—based on the national industry-occupation staffing pattern—for each OES employee in the wage records, and then we assign the most likely SOC code from the OES microdata to that employee. The rank of SOC likelihood is based on the annualized employee's wages and the public licensing data, which serves as a cross-reference check whenever an occupation requires licensure. The details on each stage mergence are explained in the program details section.

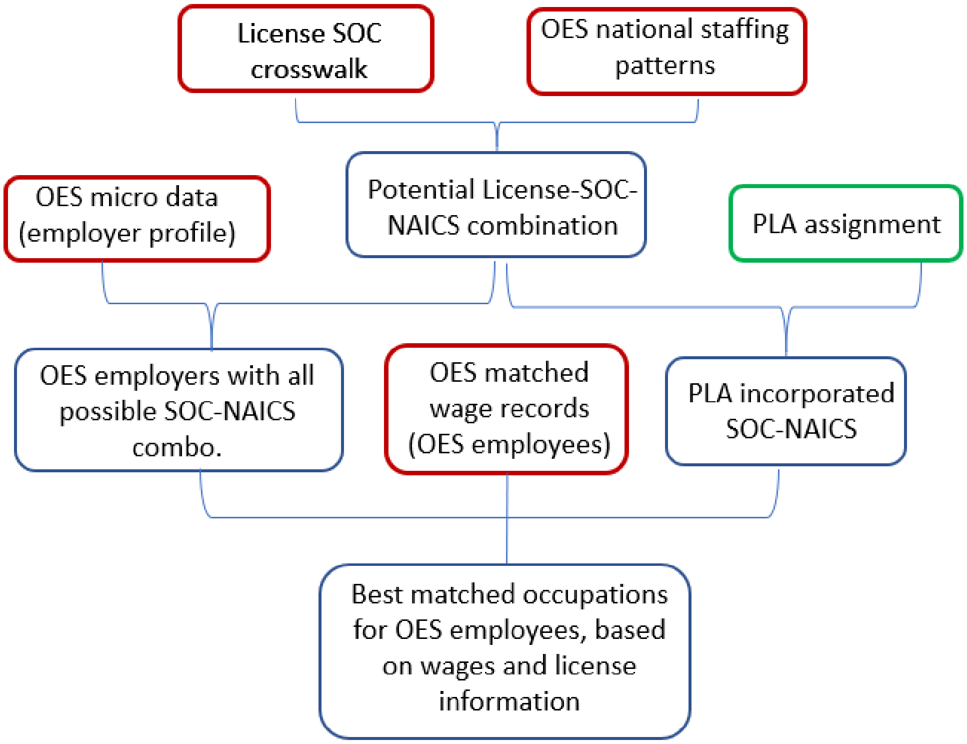

License SOC crosswalk
OES national staffing patterns
OES micro data (employer profile)
Potential License-SOC-NAICS combination
PLA assignment
OES employers with all possible SOC-NAICS combo.
OES matched wage records (OES employees)
PLA incorporated SOC-NAICS
Best matched occupations for OES employees, based on wages and license information

Figure 2: OES Microdata Assignment Process

Program Details

This section is based on "OES_micro_SOC_Assignment_dofile.do".

Step 1 – Cleaning data

This step prepares the four key data sets.

Wage records:

Data input: **OESMatchingWageRecords.csv** (provided by IWIS)

Purpose: Determine the number of jobs an employee held in a particular year and how many quarters that they held each job for within that year.

Data Clean-up:

Step 1. Annualize the wages from quarterly wages for each organization (with a unique NAICS code) an individual worked for. Take the average of the top two highest wages and annualize it by multiplying by four.

Step 2. Replace FIPS county (variable cnty) code "999" which represents missing data, and fill in the missing values with valid county codes from the same group, which can be determined by **universal_id** and **uiacct_id** jointly.

Step 3. Replace NAICS code "999999" and fill in the missing values with valid NAICS codes from the same group, determined by **universal_id** and **uiacct_id** jointly.

Step 4. Generate the wage categories that match the OES wage categories (refer to Table 3 for details).

Data output: **OESMatchingWageRecords.dta**

Description: The output data consist of universal ID (employee ID), year, UI account number (employer ID), county FIPS, 6-digit NAICS codes, working quarters, quarterly wages, number of quarters worked, number of jobs held and annualized wages and wage categories (refer to Table 3 for details), etc.

OES (wage) microdata:

Data input: **OESMicroData2013.xlsx**

Support data: **OES Microdata Field Descriptions.doc (field description)**

The input data consists of state and county FIPS code, UI account number, sample weights, 6-digit NAICS and SOC codes, and number of employees under 12 wage categories (refer to Table 3 for details).

Data Clean-up:

Step 1. Create a general ID variable and reshape the data from “wide” (employer-industry unit) to “long” form (employer-wage unit), where each wage category and industry unit of an employer makes up one record.

Step 2. Drop zero employment records, identified by the employment wage variable.

Step 3. Expand the data using employment counts to generate one record per employment. For example, if the employment count under a wage category for an employer is 2, one additional copy of such records is generated; if the employment count is 7, six additional copies of such records are generated.

Data output: **OESmicro2013_long.dta**

The output data consist of UI account number, county FIPS code, weights, industry codes, SOC codes, wage categories by industry-SOC codes, and the total number of employees of a firm.

OES national staffing patterns:

Data input: **nat4d_M2013_dl_*.xls**
(bls.gov/oes/tables.htm → download May 2013 “National industry-specific and by ownership (XLS)” zipped file)

Purpose: The raw data are separated by the national (4-digit) NAICS SOC staffing patterns. Append them into a single data.

Data Clean-up:

Step 1. Append nat4d_M2013_dl_*.xls files together.

Step 2. Drop hourly variables as well as ‘annual’.

Step 3. Create a single SOC-NAICS identifier and clean up the missing/top-coded[2] entries.

Data output: **nat4d_M2013_dl.dta**

License-SOC Crosswalk:

Data input: **license_soc_crosswalk.xlsx**

Purpose: This is an IBRC bridge file for license-SOC mapping. Prepare the information on the 2010 SOC codes associated with licenses on the license-SOC crosswalk. This will

[2] A top-coded data observation is one for which data points whose values are above an upper bound are censored. It also refers to removing erroneous outliers from the data.

indicate in the data whether a SOC code requires a license and which board oversees the licensure.

Data Clean-up:

Step 1. Drop missing SOC code (2010) observations. Drop duplicates as well.
Step 2. Create an ID variable for the SOC categories, which is then used to generate separated license profile for individual SOC codes.
Step 3. Create a "board" variable based on profession names and subsequently, generate individual board category variables based on "board" categories.
Step 4. Assign "multiple" to a "board" category for SOC codes that have multiple boards.

Data output 1: **LicenseBoard_SOC_crosswalk.dta**.

The output data consists of SOC codes and their titles, profession names, board names and each of the board categories as individual variables. This data set was used in OES SOC assignment.

Step 5. For each license SOC profile (from Step 1), repeat Step 2.
Step 6. Reshape the data from "long" to "wide", where the rows are license-SOC codes and columns NAICS codes.

Data output 2: **License_NAICS2013.dta**.

The output data consists of SOC codes and their titles, profession names, legal titles and variables for each of the 289 6-digit NAICS codes associated with licenses. This data set was also used in public licensing SOC assignment (PLA).

Licensing data:

Data input: **Licensing_IWIStoIBRC.csv (DWD file)**

Purpose: This data process works with the licensing data, and determines if a license was legally able to be used at any point during the year based on the status, date of issue, and expiration date. The input data consists of universal ID, date of birth, license type issue date, expiration date, license status, date of this status and profession name.

Data Clean-up:

Step 1. Create a new variable license rank ("licensernk") to supervise license status.
Step 2. Change the date variables with Stata date formats.
Step 3. Generate variables to indicate whether a license was active during each quarter.

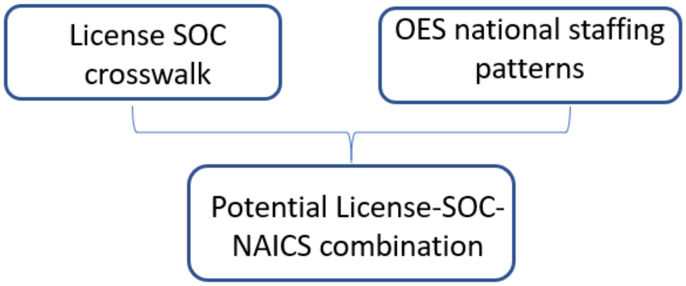
Step 4. Drop records without an active license.

Data output: **microlicensedata2013.dta**.

The output data consists of all variables in the input data and, additionally, the license active status for each quarter, ID of active license per person and total number of active licenses per person. This data set was also used in public licensing SOC assignment (PLA). For program details, refer to Licensing_SOC_Assignment.txt under the Public Licensing Agency Data module.

Step 2 – Merging data sets

Merging licensing data with NAICS-SOC codes:


License SOC crosswalk
OES national staffing patterns
Potential License-SOC-NAICS combination

Data inputs: **microlicensedata2013.dta**

Merge input: **License_NAICS2013.dta**

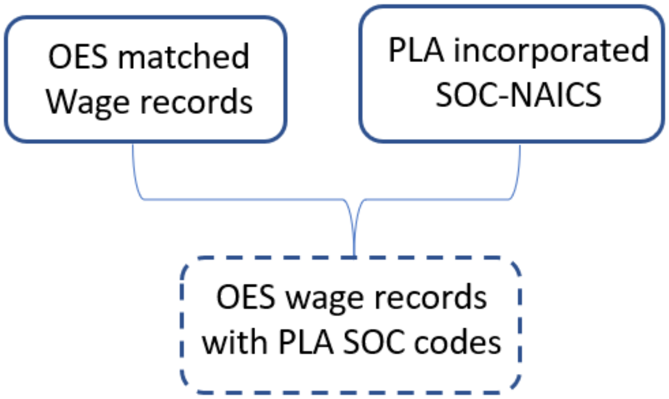
Data output 1: **LicenseOccNAICS2013.dta**

Data output 2: **PLA_byboard_2013.dta**

The first output data consists of universal ID, birthdate, license title, issuance and expiration dates, active status for each quarter, SOC codes and titles, and NAICS worked for in each quarter. This data set was also used in public licensing SOC assignment (PLA) within the Public Licensing Agency Data module.

After generating the first output, generate a board variable from the profession name variable. Aggregate licenses from the same board for the same individual into one record. This generates the second output.

Combine OES matched wage records with cleaned licensing data:


OES matched Wage records
PLA incorporated SOC-NAICS
OES wage records with PLA SOC codes

Data input: **OESMatchingWageRecords.dta**

Merge input: **PLA_byboard_2013.dta**

Step 1. After joining, generate variables to count the employment within the firm by quarter according to the wage records.

Step 2. Generate a variable that determines the lower annual wage, by taking the average of quarterly wages multiplied by the number of quarters worked.

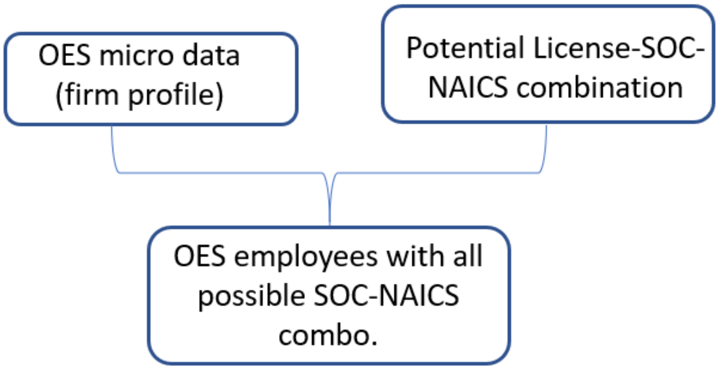
Step 3. Note that the OES matched wage records itself already has annualized *highest* wages.

Step 4. Create a match ID by combing the universal ID (individual), UI account ID (firm) and wage quarters.

Data output: **Wage_PLA_merge.dta**

The output data consists of UI account IDs, county FIPS code, year, wage quarters, 6-digit NAICS codes, quarterly wage amount, average wages of all quarters, annualized wages based on average quarterly wages and its categorization, working quarters, number of firms/jobs worked, wage ranks, annualized wage based on the highest quarterly wages and its categorization, plate status, number of months the license was held in the year, number of months the license was held in the quarter and boards.

Merging the OES microdata with the license-SOC crosswalk:


OES micro data (firm profile)
Potential License-SOC-NAICS combination
OES employees with all possible SOC-NAICS combo.

Data input: **OESmicro2013_long.dta**

Merge input: **LicenseBoard_SOC_crosswalk.dta**

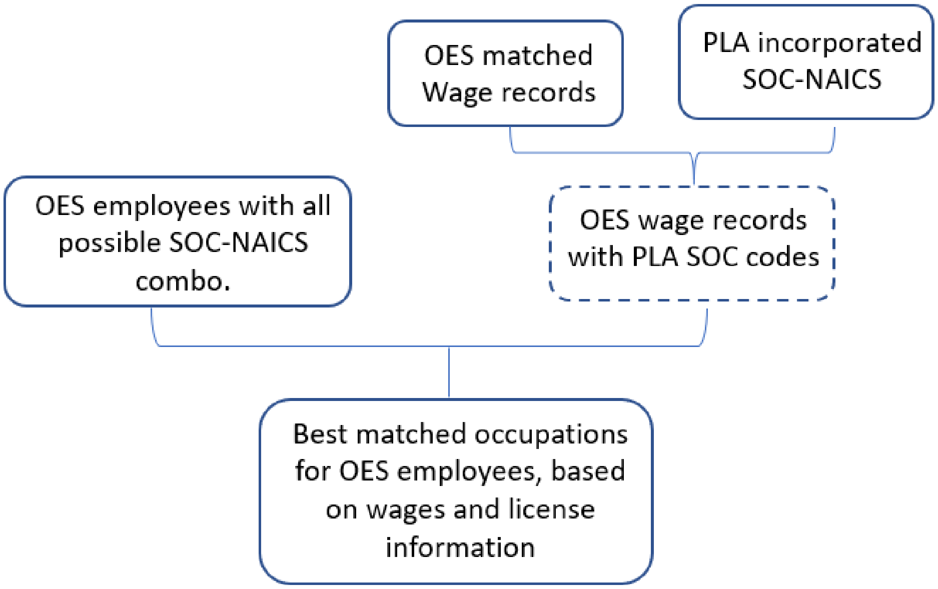
Step 1. After merging by 2010 SOC codes, generate license status according to the license-SOC crosswalk: assign license 1 if matched and 0 if unmatched.
Step 2. Join by uiacc_id with **Location counts – OES and EQUI.dta**.
Step 3. Note that the formats of UI account IDs in the two files do not match and thus must be *de-stringed* to make a match.
Step 4. Generate a variable that counts the number of employees per UI account in the OES microdata.
Step 5. Generate a variable that counts the number of different occupations per UI account (firm) in the OES microdata.

Data output: **OESmicro2013_long_wlicense.dta**

The data output consists of firm information (such as UI account IDs, legal names, EIN etc.), location counts, county FIPS code, 6-digit NAICS codes, 2- and 6-digit SOC codes, wage categories, license status and license boards.

Step 3 – Assigning SOC codes based upon best-matched wages and license information

This is a test run on NAICS 621 (health care services)


OES matched Wage records
PLA incorporated SOC-NAICS
OES employees with all possible SOC-NAICS combo.
OES wage records with PLA SOC codes
Best matched occupations for OES employees, based on wages and license information

Data input: **Wage_PLA_merge.dta**

Merge input: **OESmicro2013_long_wlicense.dta** according to UI account ID (firm).

Direct match method:

Step 1. If an individual's annual wages (either the annualized highest wages or the average annualized wages in **Wage_PLA_merge.dta**) matches with that for a potential occupation (in **OESmicro2013_long_wlicense.dta**), a SOC code is assigned to that individual.

Indirect match method:

Step 1. Create an indicator for PLA match if the licensing board for an individual in **Wage_PLA_merge.dta** matches with that in **OESmicro2-13_long_wlicense.dta**.

Note: Handling the exceptions—for professions that do not require a license under Indiana regulation—such as nurse practitioner, cosmetology, alcohol, auction and real estate—a match status of PLA is assigned to an individual whose PLA record presents one.

Step 2. Create a score based on the absolute deviations of an individual's annualized wages (from the wage records) from wages/salaries for each potential occupation.

Note: It is calculated as the sum of the lower (based on the annualized average wages) and upper (based on the highest annualized wages) bounds of the absolute deviations and is "penalized" by the PLA match status – that is, subtracting 3 for individuals whose PLA matches with that of OES microdata for firms.

Note: The higher the score, the more unlikely an individual had worked for the occupations associated with the UI account ID (firm).

Step 3. Rank the score from the smallest (most likely) to the largest (most unlikely) and remove unlikely occupations – that is, scores greater than 3 and ranks greater than 10.

Step 4. Calculate the (adjusted) total SOC and SOC-by-wages counts for each firm and use them as control margins.

Step 5. For each potential occupation under each wage category within a firm, assign the SOC code to an individual until the total number of individuals assigned reaches the control margin of SOC-by-wages counts.

Step 6. Fill in the missing observations.

Note: If an individual held the SAME active licenses for all four quarters in a year, the same SOC code is assigned for the entire year;

Note: If an individual has different quarterly SOC codes assigned with the same company, adjust them so that there is the best fit assignment for the year, and carry forward the same SOC codes for the previous quarters that were not assigned SOCs due to excessive head counts in the wage records.

Results

The results here are based on the 2013 wage records of the health care sector (NAICS 62). There are 382,679 unique person-industry (6-digit NAICS) records. The OES micro assignment assigns 352,882 cases (92 percent); public licensing (PLA) assignment fills 22,180 (6 percent) of the rest; there are still 7,617 (2 percent) cases left unassigned.

The distribution of employment among four major health care industries (3-digit NAICS)—Ambulatory Health Care Services (621), Hospitals (622), Nursing and Residential Care Facilities (623) and Social Assistance (624)—in Indiana are roughly 22 percent, 41 percent, 23 percent and 14 percent, respectively.

OES micro assignment assigns 21 major categories of occupations, of which health care practitioners, health care support, and administrative support are the three leading groups, accounting for nearly 70 percent of all occupations in health care (Figure 3). Figure 4 further shows distributional differentials by the four sub-industrial groups (3-digit NAICS). Not surprisingly, health care practitioners are the most common in ambulatory health care services (clinics) and hospitals, health care support in nursing homes, and personal care in social assistance.

Figures 5-7 present the occupational distribution for the three leading groups. Among health care practitioners, nurses—registered nurses especially—far outnumbered all other occupations combined (Figure 5). There are over 60,000 licensed nurses altogether in Indiana. Technologists and technicians are among the second most popular group, close to 30,000 combined. Therapists come in third at over 10,000. There are about 6,000 physicians and surgeons.

Health care support occupations consist of various aids to physicians, home health care and therapists, and are classified into five groups (Figure 6). There are about 20,000 more aids than nurses. Office and administrative support is the most detailed categorization, and Figure 7 presents the distribution for the top 25 categories. There are about 13,000 secretaries and administrative assistants (e.g. executive, legal and medical secretaries) in health care and the rest are various office operational clerks.

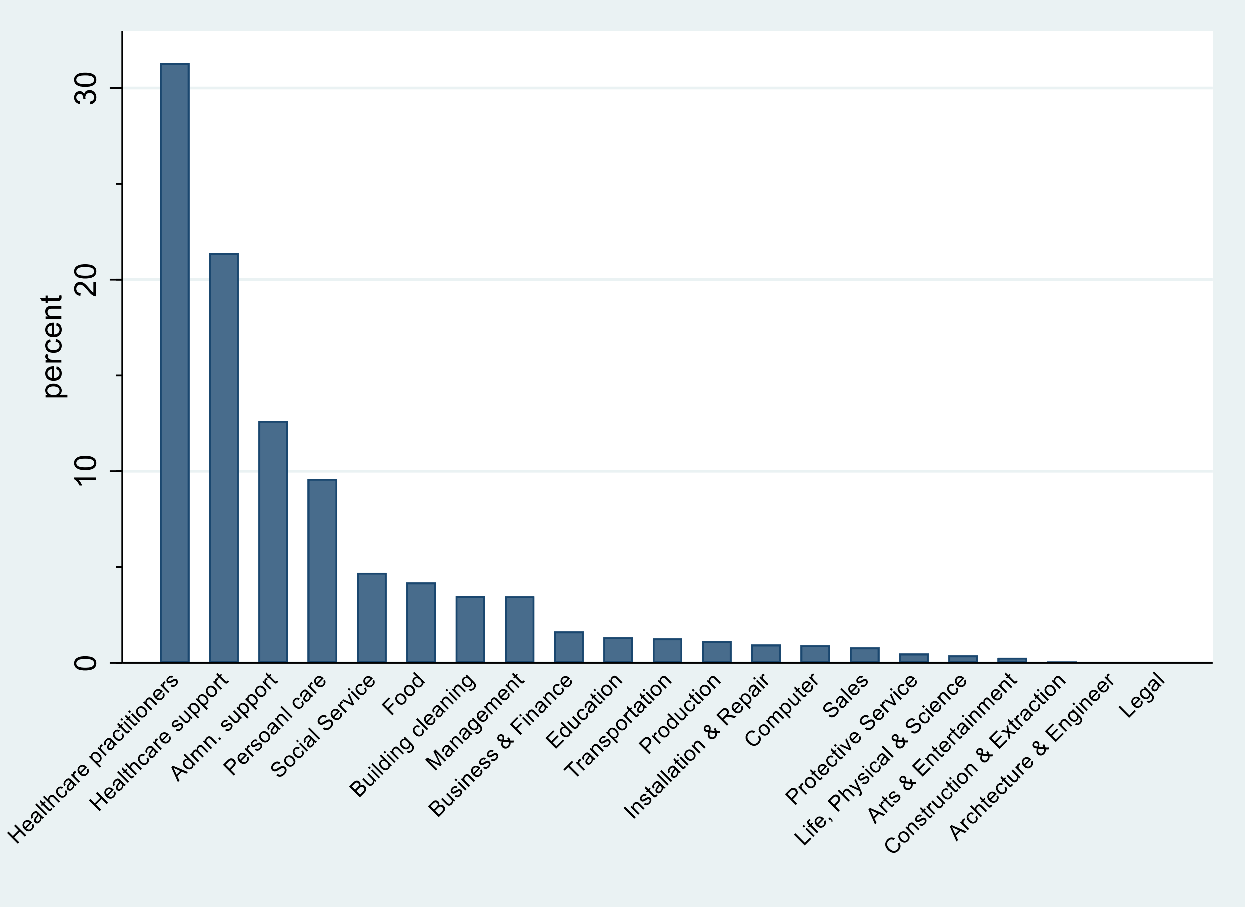


Figure 3: Distribution of major occupations in the health care sector

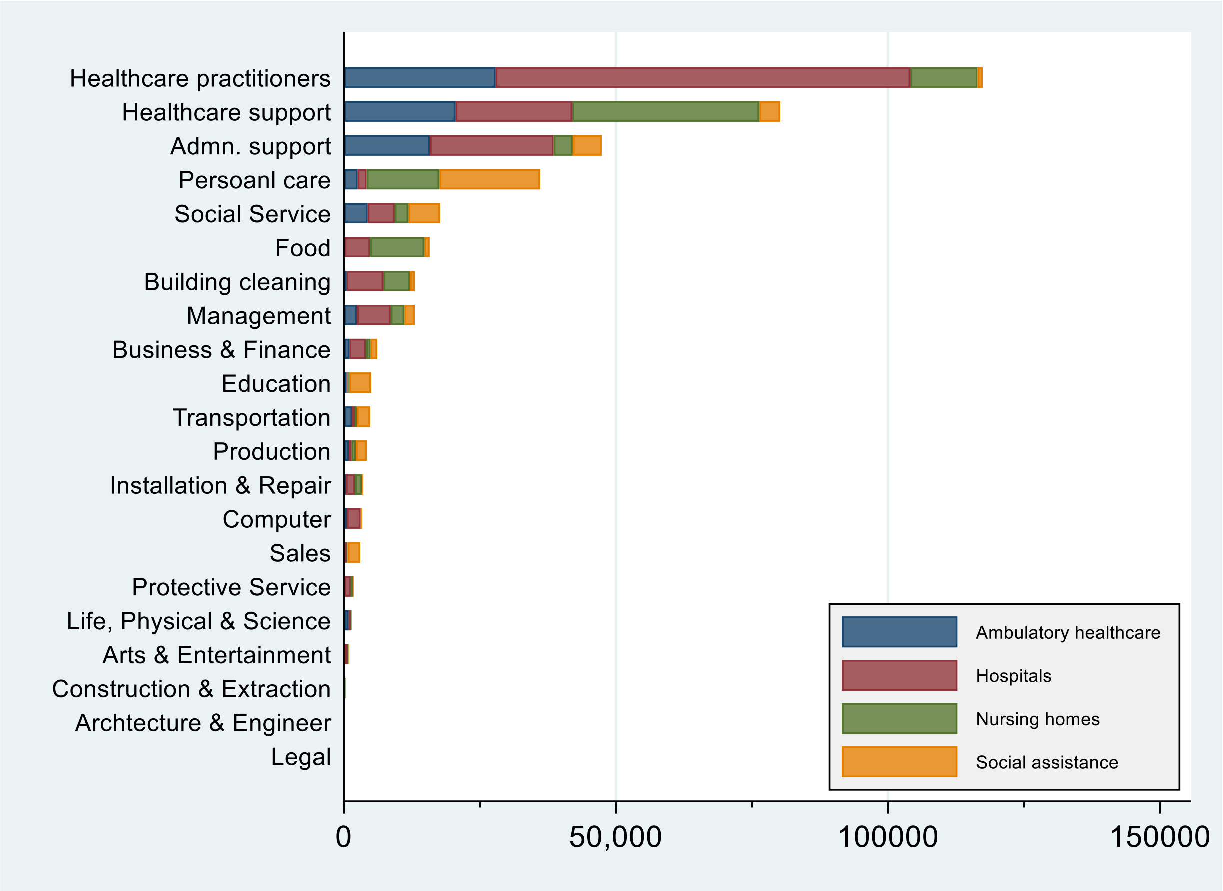


Figure 4: Distribution of major occupations in the health care sector by sub-industry

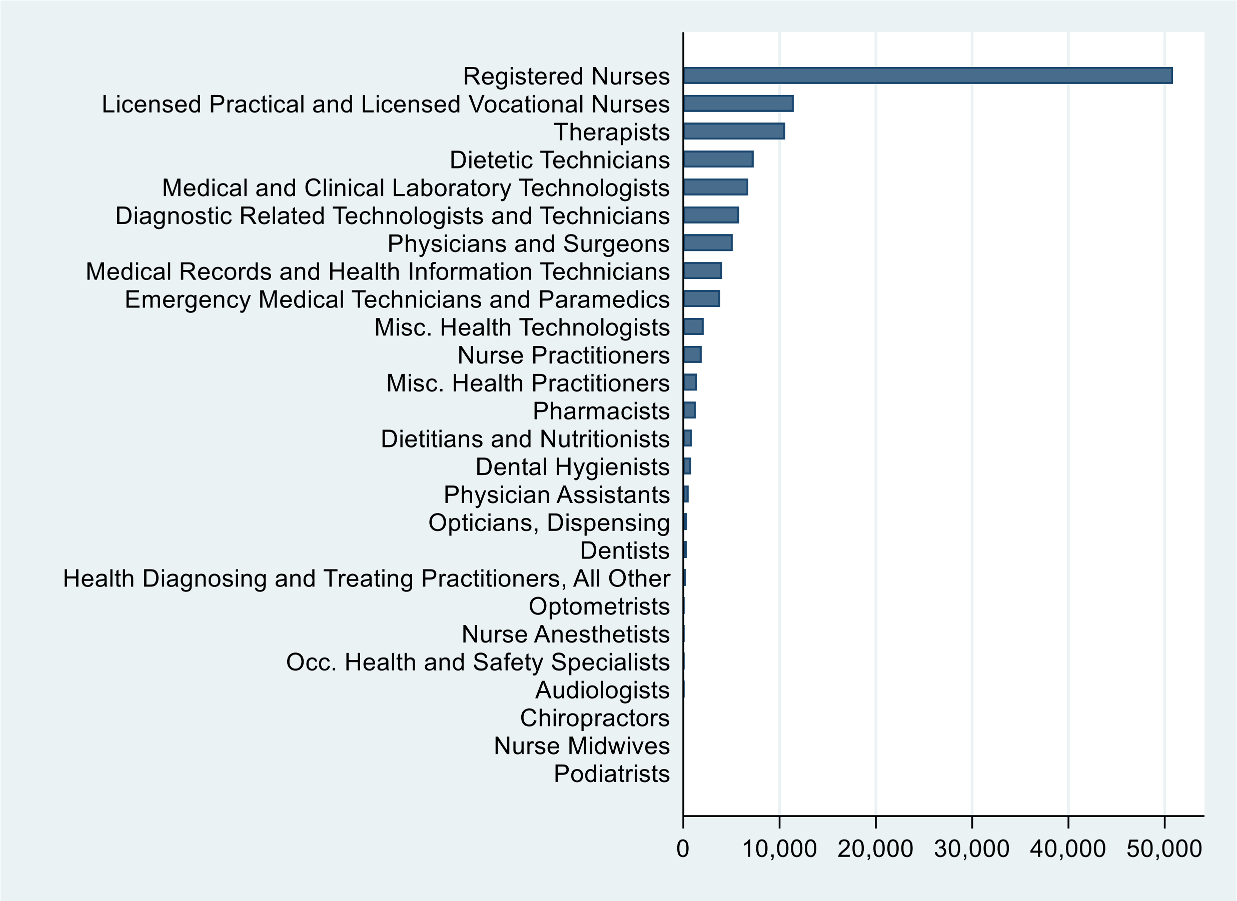


Figure 5: Occupational distribution among health care practitioners (29-0000)

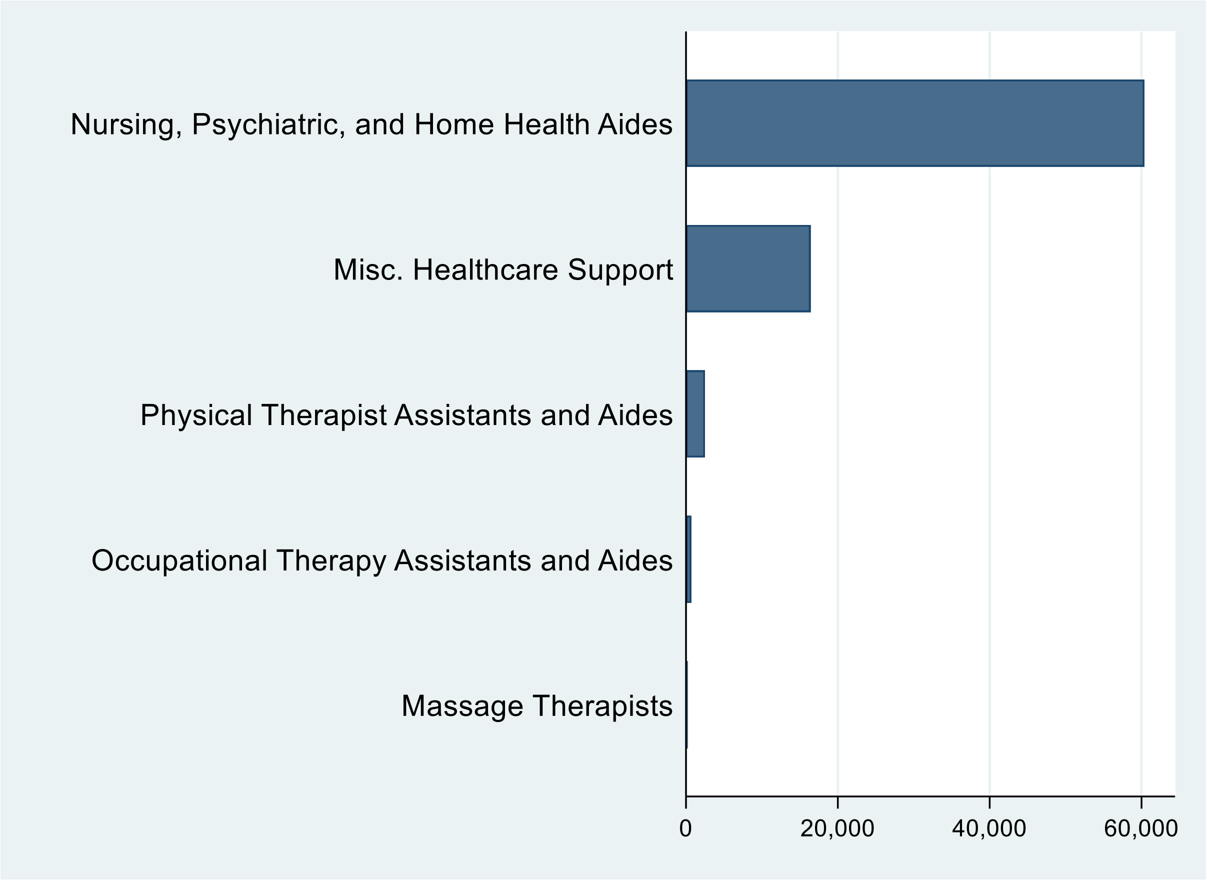


Figure 6: Occupational distribution among health care support (31-0000)

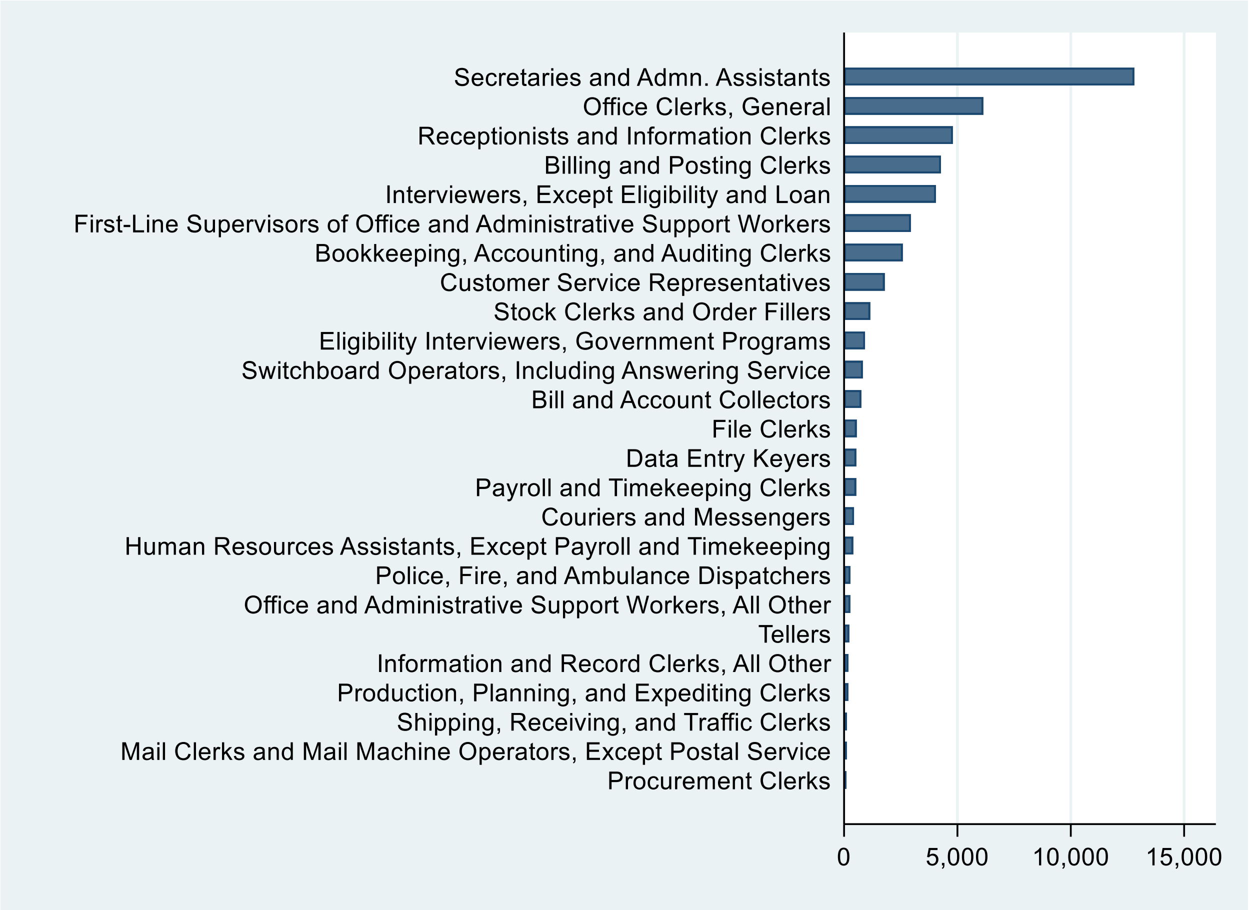


Figure 7: Distribution of top 25 occupations among administrative support (43-0000)