Imputation Methodology for the Occupational Requirements Survey (ORS) November 2016

Leland W. Righter[1], Bradley D. Rhein[1]
[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 3160, Washington, DC 20212

Abstract
The Occupational Requirements Survey (ORS) is an establishment survey conducted by the Bureau of Labor Statistics (BLS) for the Social Security Administration (SSA). The survey collects information on the vocational preparation and the cognitive and physical requirements of occupations in the U.S. economy, as well as the environmental conditions in which those occupations are performed. Imputation is a multi-step process that involves determining recipients and donors, matchmaking based on a collapse pattern and nearest neighbor, and flexible imputation based on level of and retaining collected data. This paper will describe the testing that was conducted to identify the imputation method most appropriate for generating estimates for this survey. It will also describe the imputation method that is being implemented with the estimates scheduled to be released in late 2016.

Key Words: imputation, establishment survey, imputation groups, nearest neighbor

1. Introduction

In the summer of 2012, the Social Security Administration (SSA) and the Bureau of Labor Statistics (BLS) signed an interagency agreement, which has been updated annually, to begin the process of testing the collection of data on occupations. As a result, the Occupational Requirements Survey [1] (ORS) was established as a test survey in late 2012. The goal of ORS is to collect and publish occupational information that will replace the outdated data currently used by SSA. More information on the background of ORS can be found in the next section. All ORS products will be made public for use by non-profits, employment agencies, state or federal agencies, the disability community, and other stakeholders.

An ORS interviewer attempts to collect close to 70 data elements related to the occupational requirements of a job. The following four groups of information will be collected:

- Physical demand characteristics/factors of occupations (e.g., strength, hearing, or stooping)
- Specific vocational preparation requirements, which include educational requirements, experience, licensing, and certification and post-employment training
- Mental and cognitive demands of work
- Environmental conditions in which the work is completed

The survey plans to publish all estimates that meet the reliability and confidentiality criteria. Somewhere between three and eighteen estimates will be calculated for each of the 70 ORS data elements. Around 920 total estimates could be calculated for a single occupation or occupational group. Estimate types include the percentage of workers in a given category, mean, percentiles (10%, 25%, 50%, 75%, and 90%), and the mode.

This paper explores the ORS imputation processes. Item non-response, within a given occupational quote, is the target of the imputation process. If a respondent did not provide or was unable to provide details relative to a certain occupation then this imputation process would supply the missing details. Section 2 provides background information on the Occupational Requirements Survey. Section 3 summarizes the ORS data elements and types of estimates that will be calculated for each. Section 4 details the original specifications of the ORS imputation process, including what the logic was relative to the initial design of the survey. Section 5 outlines the research process for the ORS and why modifications were made to the design. Section 6 details the specifications for the final ORS imputation design.

2. Background Information on ORS

In addition to providing Social Security benefits to retirees and survivors, the Social Security Administration (SSA) administers two large disability programs, which provide benefit payments to millions of beneficiaries each year. Determinations for adult disability applicants are based on a five-step process that evaluates the capabilities of workers, the requirements of their past work, and their ability to perform other work in the U.S. economy. In some cases, if an applicant is denied disability benefits, SSA policy requires adjudicators to document the decision by citing examples of jobs the claimant can still perform despite restrictions (such as limited ability to balance, stand, or carry objects) [2].

For over 50 years, the Social Security Administration has turned to the Department of Labor's Dictionary of Occupational Titles (DOT) [3] as its primary source of occupational information to process the disability claims [4]. SSA has incorporated many DOT conventions into their disability regulations. However, the DOT was last updated in its entirety in the late 1970's, although a partial update was completed in 1991. Consequently, the SSA adjudicators who make the disability decisions must continue to refer to an increasingly outdated resource because it remains the most compatible with their statutory mandate and is the best source of data at this time.

When an applicant is denied SSA benefits, SSA must sometimes document the decision by citing examples of jobs that the claimant can still perform, despite their functional limitations. However, since the DOT has not been updated for so long, there are some jobs in the American economy that are not even represented in the DOT, and other jobs, in fact many often-cited jobs, no longer exist in large numbers in the American economy.

SSA has investigated numerous alternative data sources for the DOT, such as adapting the Employment and Training Administration's Occupational Information Network (O*NET) [5], using the BLS Occupational Employment Statistics program (OES) [6], and developing their own survey. SSA was not successful with any of these potential data sources and turned to the National Compensation Survey [7] program at the Bureau of Labor Statistics.

3. ORS Data Elements and Possible Estimates Summary

The ORS is designed to capture occupational information on educational requirements, cognitive and physical demands, and exposures to environmental conditions. An extensive description of ORS data elements and how estimates for each element will be calculated can be found in the paper "Estimation Considerations for the Occupational Requirements Survey" [8]. Information on estimation processing can be found in the paper "Estimation Processes Used in the Occupational Requirements Survey" [9].

Many of the ORS data elements will have the percentage of workers, mean, percentiles, and modes estimates for each occupational definition. For example, one ORS data element measures the amount of time during a typical day that a worker, such as a nurse, spends stooping. Occupational definitions are derived from the Standard Occupational Classification Manual (SOC) [10] and O*NET. Physical demands, such as stooping, are captured in hours and are also converted to percent of the day, and so mean and percentile estimates (10%, 25%, 50%, 75%, and 90%) will be calculated for both hours and percent of the day. Also, the hours of time spent stooping will fall within an SSA-established category, and so a percentage of workers estimate will be calculated for each category. SSA defines five categories by a range of hours spent performing an activity – not present, seldomly, occasionally, frequently, and constantly. Finally, the mode of the categories will be identified, marking the eighteenth estimate related to stooping.

4. Original Specifications of the ORS Imputation

Before reading the next section, some vocabulary needs to be defined. A "donor" is a known data element available to be imputed for unknown data. A "recipient" is a missing data element that requires the known data element of a "donor" in order to become known data.

The original idea for an ORS imputation design was a nearest neighbor method based on the employment of establishments connected with the respective occupations. For unknown data at the item level, a search would be performed to find an occupation with known data from the establishment with the closest employment size (an absolute difference of 0 is the most desirable). Beyond the usage of nearest neighbor, a thirteen element collapse pattern would be used to match occupations of known and unknown data. If a match could not be made for all thirteen elements, an element would be removed and a match would be attempted for twelve elements instead. This process would continue until the two non-collapsible elements, deemed essential to match on, were reached. Table 1 displays the original collapse pattern,

Table 1: Original Collapse Pattern of the ORS (Non-collapsible elements highlighted in yellow. First collapsed element is order number 13, second is 12, etc.) (For definitions of variables see Appendix A)

Order	Cell Variable Name
1	Ownership
2	2-digit SOC
3	3-digit SOC
4	5-digit SOC
5	6-digit SOC
6	8-digit SOC
7	Supervisory status
8	Major Industry Division
9	Two-digit NAICS code
10	Establishment Size Class
11	Union/Non-union status
12	Full-time/part-time status
13	Census region

Donor usages were kept track of during the entirety of the imputation process. While the collapse pattern was being exhausted of potential matches, donors were not to be used more than 3 times, and a donor with a usage less than another donor would get priority in the matching regardless of employment difference. For example: A donor has been used once and has an employment difference of 1000, while another donor has been used twice and has an employment difference of 50; the donor used only once would get priority and would be matched. If the entire collapse pattern was exhausted and there were still recipients with unknown data, then the limit of 3 donor usages would be eliminated. Imputation would start at the full collapse pattern (matching for all 13 elements) again and would still use previous donor usages as a factor in priority.

Because the diverse variables in the ORS can be categorized into logically-related groups, we concluded that group imputation might work effectively. The original design outlined 13 imputation groups, consisting of anywhere from one to tens of variables. These 13 groups are defined in table 2. When imputing in groups, all data in a recipient's group was replaced by donor data; for example, if 3 out of 5 variables had been collected in the group, meaning that 2 were unknown, then the entire group would be imputed and even the collected data would be overwritten by imputed data. The idea behind this was to maintain consistency in our data and ensure that our imputation process preserved the relationship among variables.

Table 2: Original Imputation Groups

Group	Label
1	Cognitive
2	Driving
3	Vision
4	Hearing
5	Environmental Conditions
6	Climbing
7	Postural
8	Keyboarding & manipulation
9	Legs/Feet
10	Arms/Hands
11	Standing
12	Lifting
13	SVP

4. ORS Imputation Research

Research for the ORS imputation process initially progressed based off of the original design using pre-production survey data. The effectiveness of the imputation process was measured by a number of factors. The first factor was the magnitude of changes in the estimation process, what we dubbed "Big Changes"—we would run the collected data through estimation, then run the collected data plus the imputed data through estimation, and compare the two resulting estimate files. Ideally, the values should not drastically change in imputation without a reasonable explanation. The next factor used to determine effectiveness was the number of recipients with unknown data that remained after imputation. Also, run time and variances were used to assess how well the imputation process was working.

While the number of remaining recipients was favorable with the initial design, there were concerns over the number of Big Changes. A series of tests followed where reinforcements were made to the collapse pattern and the imputation groups. Some of the reinforcements included: removing variables from the collapse pattern, changing the order of the variable "size class" in the collapse pattern, redefining the classes of "size class." Unfortunately these adjustments did not lower the number of Big Changes to an acceptable level.

In order to reduce the number of Big Changes we decided that a technical modification might be necessary to attain desirable results. A hybrid of nearest neighbor imputation and cell means imputation was proposed, and tests were carried out using these methods together. For this process the nearest neighbor technique would find donors for non-numerical data, while the cell means technique would find donors for quantitative data. The cell means technique collects a pool of donors based on matches in the collapse pattern and then derives an overall mean from the pool. After the mean has been calculated it is

donated to a group of recipients with matching components of the collapse pattern. While the number of Big Changes did decrease for this method, there were concerns that the variances of our estimates would be artificially low, given that the imputed values for cell means all have the same value.

The combination of cell means and nearest neighbor as an imputation design was deemed unfavorable, so research again focused on a pure nearest neighbor design. After taking a closer look at the nature of the data (and becoming more familiar with it), technical modifications were proposed that would affect details of the group imputation along with adding more layers to the imputation process.

One major concern thus far in the research had been the overwriting of collected data in the imputation groups, and it was theorized that this may have been affecting the Big Changes we were seeing. In order to maintain consistency and not lose any collected data, a Collected Data Retention (CDR) process was implemented. If, for example, a group had 3 out of 5 variables collected (the other 2 missing) then, with the CDR, imputation would be performed for those 2 missing variables and the three collected variables would not be overwritten by donor data. This totally eliminated any loss of data and kept the ideology behind group imputation intact.

A complexity of the data also inspired another technical change. When asking about the presence of certain variables, like reaching overhead, it was allowed for the respondent to say, "I know that the employee reaches overhead, I just don't know how long they have to do it for." This was called "Present Duration Unknown." What this meant was that if a group had Present Duration Unknown as a value, then the donor would have to have a positive duration for the associated variable. For example: If reaching overhead is present duration unknown, then the donor would need to have reaching overhead duration greater than 0 to ensure consistency of data. This stipulation would limit the pool of donors for this potential imputation group, which might adversely affect the final result. The solution to this was multilayered imputation. Instead of simultaneously imputing for reaching overhead presence and reaching overhead duration (and doing this for all variables in the imputation group), the auxiliary variables (variables like duration) are separated and a different layer of imputation would be performed.

Here's an example of multilayered imputation:

- Layer 1—the presence of reaching overhead and a number of other variables in an imputation group are unknown. Imputation is performed to designate if these variables are present or not.
- Layer 2—It has been determined that reaching overhead is present for this occupation, but there is an unknown value for the duration of reaching overhead. Imputation is performed to find a value for the duration. (If reaching overhead was not present then no imputation would be performed.)
- Layer 3 and every proceeding layer—Imputation is performed or not performed based on the previous results of other layers of imputation.

Part of the imputation process is instituting constraints in the matchmaking process so the pool of potential donors is small enough to adequately match for the recipients and large enough that every recipient will find a match. For the imputation design of the ORS, there were 3 constraints put into place for recipient-donor matches: employment difference, the variables of the collapse pattern, and the known/unknown status of the variables of the imputation groups. After discussions and testing, it was determined that the employment

difference and the collapse pattern variables were the most important factors in matching. So in order to open up the pool of donors to reduce the number of recipients without matches, it was decided that modifications to the imputation groups would be the most effective change. By breaking down the groups into layers, the number of potential donors broadened at each layer, and the number of recipients without matches diminished.

5. Specifications of Final Design for Imputation

The original technique of using nearest neighbor to match recipients and donors based on employment was utilized in the final design. Along with the additions of the CDR and multilayered imputation, changes were made to the collapse pattern and the number of imputation groups. The reason for these changes was to further reduce the recipients without donors (by adding a broader element to the collapse pattern and reducing the non-collapsible elements) and to increase our pool of donors for each imputation group (by dividing some of the larger groups).

In order to reduce the recipients without donors, we introduced a variable called "Broad SOC." SOC (Standard Occupational Classification) is a numerical system identifying and categorizing a number of jobs. With each unique SOC number assigned to a job, there is a hierarchy within the SOC number that includes similar jobs. The more digits included, the more exact the job listing is. So a 2 digit SOC might identify all construction and extraction occupations, while a 6 digit SOC will identify a specific type of miner. With the Broad SOC we created a new categorization combining some of the 2 digit SOCs so that a broader pool of donors could be created. The new collapse pattern can be found in table 3 and the new imputation groups can be found in table 4. The limitation for donor usage remained at 3 for the final design, and the limitation would be lifted if there were still recipients without donors.

Table 3: Final Collapse Pattern of the ORS (Non-collapsible elements highlighted in yellow. First collapsed element is number 10, second is 9, etc.)

Order	Cell Variable Name
1	Broad SOC
2	Ownership
3	2-digit SOC
4	3-digit SOC
5	5-digit SOC
6	6-digit SOC
7	8-digit SOC
8	Two-digit NAICS code
3	Establishment Size Class
9	Union/Non-union status
10	Full-time/part-time status

Table 4: Final Imputation Groups (Groups that were formerly conjoined and are now split in two are highlighted.)

Group	Label
1	Cognitive
2	Driving
3	Vision
4	Hearing
5	Weather
6	Environmental conditions
7	Climbing
8	Postural
9	Keyboarding
10	Manipulation
11	Legs/Feet
12	Arms/Hands
13	Standing
14	Lifting
15	SVP

7. Conclusion and Ongoing Research

This research shows that imputation should be tailored to fit the unique properties of a particular survey's data. The changes from the initial design were minimal, more like adjustments made to optimize the additional technical changes (the CDR and multilayered imputation), which were much more integral to the effectiveness of imputation and ensuring that there were no radical changes to the estimates.

Going forward, as ORS data elements change and BLS starts to build a larger sample, research will continue to study the interaction between the established imputation design and the data, and adjustments will be made as more becomes known about that interaction.

References/Footnotes

[1] Occupational Requirements Survey, http://www.bls.gov/ors/.
[2] Social Security Administration, Occupational Information System Project, http://www.ssa.gov/disabilityresearch/occupational_info_systems.html.
[3] U.S. Department of Labor, Employment and Training Administration (1991), "Dictionary of Occupational Titles, Fourth Edition, Revised 1991"
[4] Occupational Information Development Advisory Panel, 2010, http://www.socialsecurity.gov/oidap/index.htm
[5] U.S. Department of Labor, O*Net Online, http://www.onetonline.org/
[6] Bureau of Labor Statistics, Occupational Employment Statistics Program, http://www.bls.gov/oes/
[7] National Compensation Survey, http://www.bls.gov/ncs/.

[8] Rhein, Brad, Ponikowski, Chester, and McNulty, Erin. 2014. Estimation Considerations for the Occupational Requirements Survey. In *JSM proceedings*, Government Statistics Section. Alexandria, VA: American Statistical Association. 2134-2146.

[9] Rhein, Bradley D., Ponikowski, Chester H. 2015. Estimation Processes Used in the Occupational Requirements Survey. In *JSM proceedings,* Government Statistics Section. Alexandria, VA: American Statistical Association.

[10] Standard Occupational Classification System, http://www.bls.gov/soc/.

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics or the Social Security Administration.

Appendix A – Collapse Pattern Variable Definitions:

Ownership: Designates if the ownership is Private, Civilian, or Government.

SOC (Standard Occupational Classification): A standardized numerical classification system for occupations, used with each job within a surveyed company for an assigned title.

Supervisory Status: Designates if the occupation is supervisory or not.

NAICS (North American Industry Classification System): A standardized numerical classification system used to identify the industry of surveyed companies.

Major Industry Division: Designates the major industry of each company, defined by 2-digit NAICS code.

Establishment Size Class: A classification system that assigns each establishment with an employment size.

Size Class	Employment Range
1	0-49
2	50-99
3	100-499
4	500+

Union/non-union status: Designates if the occupation is union or non-union.

Full-time/part-time status: Designates if the occupation is full or part time.

Census Region: A numeric system used to identify what region each company does business in.

Census Region Code	Census Region Name	States Included
1	Northeast	ME, CT, MA, NH, RI, VT, NJ, NY, PA
2	South	AL, KY, MS, TN, DE, DC, FL, GA, MD, NC, SC, VA, WV, AK, LA, OK, TX
3	Midwest	IL, IN, IA, MI, WI, OH, KS, MN, MO, NE, ND, SD
4	West	AZ, CO, ID, MT, NV, NM, UT, WY, AK, CA, HI, OR, WA