

# Effect of Nearest Neighbor Imputation on Variances Calculated by Fay's Balanced Repeated Replication

Bradley D. Rhein<sup>1</sup>, Chester H. Ponikowski<sup>1</sup>, and Leland Righter<sup>1</sup>

<sup>1</sup>U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 3160,  
Washington, DC 20212

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference

## 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 70 data elements that cover 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. Since some sample units are not willing or able to provide data for all data elements, missing data element values are imputed using a nearest neighbor imputation procedure. In cases where there are multiple eligible donors, a random selection process selects a donor. Variance estimates are generated after imputation using the Fay's Balanced Repeated Replication method. Since imputation runs on the full sample before variance estimation, and not for each replicate during variance estimation, the variances are deemed to be underestimated. This paper presents the research results for comparing the current method with a method where imputation occurs at each replicate.

## 1. Introduction

In the Occupational Requirements Survey (ORS), the current estimation process uses both imputed and collected values in the calculation of variances. Missing items are imputed using a nearest neighbor imputation method. In cases of ties, that is, where there are multiple eligible donors, a random selection process is used to select a donor. . Variances are calculated using Fay's Balanced Repeated Replication, where all sampled occupations (quotes) contribute to each sample stratum-based replicate estimate after a weight adjustment.

A concern is that the ORS variances are underestimated since the variance calculation does not account for the random process used to select a donor in the case of a tie during imputation, and around 10 percent of the donor to recipient matches are decided by a random process.

The purpose of this paper is to analyze these estimated variances for ORS. The next section gives a brief background of the survey, followed by brief descriptions of the sampling, imputation, and estimation processing for ORS. Then, there's a discussion of the research question, followed by an empirical evaluation of the current and proposed methods. Results are then summarized.

## 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. A final determination about which citizens, or claimants, are eligible to receive benefits are based on a five step process that evaluates the capabilities of the worker, the requirements of their past work (prior job), and their ability to perform work for any job in the U.S. economy. 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 their restrictions (such as limited ability to balance, stand, or carry objects) [1].

For over 50 years, the Social Security Administration has turned to the Department of Labor's Dictionary of Occupational Titles (DOT) [2] as its primary source of occupational information to process the disability claims [3]. 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 available data at this time.

The purpose of the ORS is to collect the various physical demands, environmental conditions, education and training, and mental requirements for occupations within the national economy. The information in ORS is unique, compared with other job requirement documentation (such as the DOT or the Occupational Information Network (O\*NET) [4]), as the data give a better understanding of some of the cognitive and mental requirements for a job. In addition, ORS provides insight into the duration of specific physical demands and environmental exposures, as well as the amount of education, training, and experience needed to perform in the occupation.

ORS data are reported on the website [5] as news releases, database tables for query, infographics, and occupational profiles.

When an applicant is denied SSA benefits, SSA documents the decision by citing examples of jobs that the claimant can still perform. But some of the jobs in the American economy are not even represented in the DOT and other jobs, in fact many often cited jobs, don't exist in large numbers in the American economy any longer. For example, a job that is often on the list for applicants is "envelope addressor." If this job still exists in our economy, there aren't too many of them and the positions are hard to find.

### **3. Overview of the ORS Sampling, Imputation, Estimation, and Variance Estimation Methods**

#### *Sample Design*

The ORS sample design is a two-stage stratified national sample of establishments and the occupations within those establishments. The establishment sampling frame was developed from the BLS Quarterly Census of Employment and Wages (QCEW) database [6] with the addition of railroads that are not included in the QCEW. Stratification of the sampling frame is by industry and ownership, directly, and also implicitly by region. Private industry and State and local government establishments are included, and industries are defined by the North American Industry Classification System (NAICS) [7].

The allocation of establishments to sampling strata is proportional to the stratum establishment employment size. At the first stage of sampling, establishments are selected from each stratum by systematic probability proportional to establishment employment size (PPS) sampling. At the second stage, occupations (quotes) are then sampled from the selected establishments by a probability proportional to occupation employment size procedure and classified by the Standard Occupational Classification (SOC) [8] and 8-digit SOC codes provided by O\*NET.

ORS samples currently follow a three-year rotation. For more details on the current sample design, see "Occupational Requirements Survey Sample Design" by Ferguson, et al. [9].

#### *Imputation Method – Nearest Neighbor with a random hot deck component*

In ORS, missing items occur at the quote level. These missing items are imputed by a nearest neighbor procedure within a defined, and collapsible, imputation cell. Nearest neighbors are determined by the establishment employment size. In the event that there are multiple nearest neighbors, a random process decides which donor will be used.

An occupation with collected data that is the "nearest neighbor" may donate data to an occupation that has missing items. A donor is the "nearest neighbor" because it shares many characteristics (noted in Figure 1 below) with the occupation that has missing items. The assumption is that items are missing at random.

Only actual data collected in the field can be donated to a recipient. Items that have been shown to be related, such as the presence of fine and gross manipulation, are imputed in groups to keep the relationships intact. Certain items' relationships are maintained by completing imputation in a specific order. For example, a physical demand must be

imputed as present before a duration could be imputed. Other items, such as the duration of activities, are imputed individually since they have been found to have no significant correlation.

Characteristics shared by occupations form the imputation cell, defined by SOC groups, ownership, specific SOC codes, industry groups, size class, union status, and full/part time status. These characteristics, all of which must be collected for a quote to be considered available for imputation, are captured for each quote used in the estimation process. Since ORS is a survey of occupations, imputation occurs among quotes that share as many occupational characteristics as possible. The most important characteristic, then, is the 8-digit SOC code from O\*NET.

The current full imputation cell appears below in Figure 1. If there are no available donors in the full imputation cell, the process removes a characteristic, beginning at the end of the list, and continues to search for a donor. Characteristics may continue to be removed to find a donor, however, the first characteristic, broad SOC, must always remain.

Figure 1: Imputation Cell Definition

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
9	Establishment Size Class
10	Union/Non-union status
11	Full-time/part-time status

Finding the “best” donor-to-recipient match requires two steps. First, all recipients are matched to their respective nearest neighbor - the closest donor in employment and within imputation cell. After each use, a donor loses priority to the other donors that have been used fewer times. If multiple donors are deemed the nearest neighbor, a random process determines which donor is the nearest neighbor.

Donors may be used up to three times, unless there is a severe lack of data for a particular data element. In these circumstances, a donor may be used more than three times.

For a detailed explanation of the ORS imputation method used through November, 2017, see “Imputation Methodology for the Occupational Requirements Survey” by Richter et al. [10]

*Estimation*

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” [11]. Information on estimation processing can be found in the paper “Estimation Processes Used in the Occupational Requirements Survey” [12].

The ORS estimates include the percentage of workers, mean, percentiles, and mode 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, as defined earlier). Physical demands, such as reaching, 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%) are calculated for both hours and percent of the day. Also, the hours of time spent reaching fall within an SSA-established category, and so a percentage of workers estimate is calculated for each category. SSA defines five categories by a range of hours spent performing an

activity – not present, seldom, occasional, frequent, and constant. Finally, the mode of the categories is identified, marking the eighteenth estimate related to reaching.

#### *Variance Estimation*

Variance estimates are calculated using Fay’s Balanced Repeated Replication (BRR), as detailed in “Variance Estimation for the Occupational Requirements Survey” by Rhein, Ponikowski [13]. Fay’s version of BRR allows all observations in the sample to appear in each replicate half-sample, albeit with a weight adjustment. Using a method that allows all quotes to contribute to each replicate aids in estimating the variances because the variance strata for some occupations depend on a relatively small amount of quote-level data. The ORS uses 236 variance strata.

Replicate half-samples of the original sample within each variance stratum are constructed by assigning, at random, half of sample establishments to half-sample 1 and the other to half-sample 2. Individual replicates are formed according to a pattern of “+1” and “-1” symbols that are found in the rows of 236 by 236 Hadamard Matrix. The “+1” indicates that half-sample 1 units are selected and “-1” indicates half-sample 2 units are selected within a given stratum. Then, occupational replicate weights are increased or decreased depending on the half-sample selection flag. All usable quotes were randomly assigned a half-sample selection flag during sampling and will appear in each replicate half-sample with an appropriate weight adjustment.

Once the replicate half-samples are established and the occupational replicate weights have been adjusted, estimation is run for each of the replicate half-samples. Variances are then calculated for each estimate, using the sum of the differences between the full sample estimate and each of the replicate estimates. Full sample estimates were calculated using the final occupational weights.

#### **4. Overview of the Research Question**

This paper attempts to clarify whether the uncertainty inherent to the imputation process is appropriately accounted for in the variance estimation procedure. Currently, the randomness inherent in imputation is not accounted for in the variance calculation. The hypothesis, then, is that the variances are underestimated.

Under the current process, ORS data receives establishment and occupational-level weight adjustments before missing items are imputed by a nearest neighbor, hot deck imputation method where ties of donor-to-recipient matches are broken randomly. After imputation, both the collected and imputed data values contribute to estimation and variance estimation.

The imputation method is not currently repeated during variance estimation. As a result, the calculation of variances does not account for the randomness associated with the imputation method. Depending on the group of data elements (listed below), the prevalence of randomly assigned donors is between 10% and 14%.

*Figure 2: Percentage of donor to recipient matches determined by ties, by group of ORS data elements*

Group Label	Percentage of matches determined by ties
Driving	11.42%
Vision	14.34%
Hearing/speaking	13.62%
Heat/cold/humid	14.23%
Hazardous	13.86%
Climbing	14.94%
Postural	13.84%
Keyboarding	15.00%
Manipulation	10.05%
Pushing Legs	13.84%
Pushing Arms	13.77%
Sitting/Standing	11.43%
Lift/carry	13.86%
Education	11.39%

Ninety percent of all ties have been found to occur as the result of a single establishment containing more than one job with the same SOC code. For example, an establishment could have the following list of jobs, shown in Figure 3 (note that the data is fake). Note that three of the jobs, while having slightly different job titles, all have the same SOC code. Note that jobs 3, 7, and 8 are the exact same as Jobs 2, 4, and 5, respectively.

*Figure 3: Job List Example*

<u>Job #</u>	<u>SOC</u>	<u>Title</u>	<u>Full/PartTime</u>	<u>Salary/Incentive</u>
1	119000	Manager	Full	Salary
2	412000	Sales Rep	Part	Incentive
3		2		
4	499000	Maintenance	Full	Salary
5	412000	Senior Sales Rep	Full	Salary
6	412000	Sals Rep	Part	Incentive
7		4		
8		5		

Now, the procedure for collecting occupations allows for multiple SOC codes to be collected if sampled, and all sampling of occupations is by a probability proportional to employment size procedure. Usually, if more than one job within the same SOC code is sampled, one sampled SOC will be given twice the occupational weight (i.e. Job 2 has twice the weight because Job 3 is the same as Job 2). The goal, then, is to collect truly unique occupations within an establishment.

However, occupations with different full/part time status, union/non-union status, or salary/incentive pay status would be considered unique even if the SOC code is the same. In ORS, there could be many different combinations of “unique” occupations. Some of these “unique” occupations are more unique than others. For example, in Figure 3 above, Jobs 2 and 6 are not discernably unique.

Two methods have been employed to study the effects of the randomness inherent in the imputation method on variances:

1. Current method – run imputation once, before variance estimation, and calculate variances by replication using both the collected and the one set of imputed values
2. Proposed method – run imputation once for each replicate while calculating the variances, using both collected and imputed values, where the imputed values vary for each replicate half-sample

Running imputation for each replicate (method #2) accounts for the randomness inherent in the imputation method. Such a method was proposed in a paper by Andridge and Little titled “A Review of Hot Deck Imputation for Survey Non-response” [14].

Both methods use the same variance calculation (Fay’s BRR), including the same variance strata, panel assignments, Hadamard matrix, and replicate weights. The next section will detail the analysis of the empirical results.

## 5. Analysis of the Empirical Results

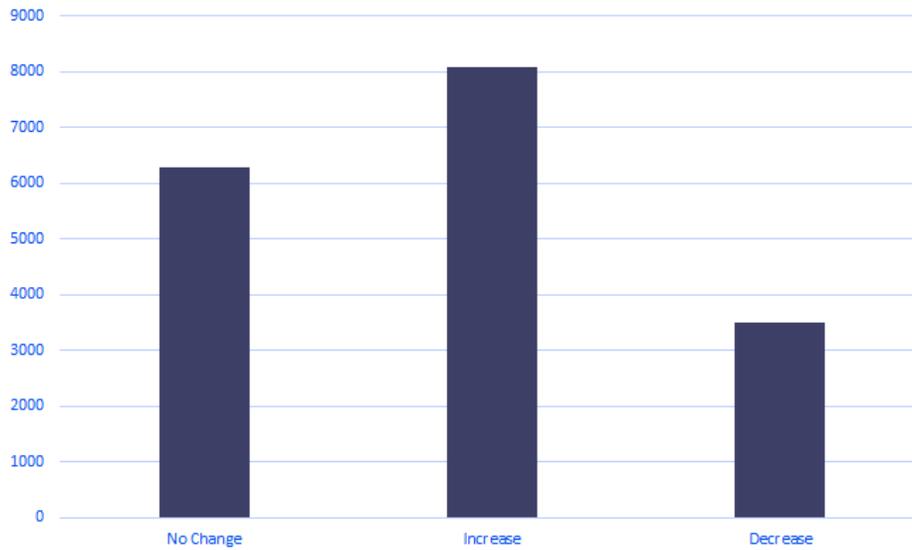
In order to test the two methods, a smaller set of estimates were studied. The data elements were chosen in such a way that all types of ORS data elements were represented, including data elements with characteristics such as binary responses, mean and percentile calculations, and data elements having to do with physical demands, environmental conditions, and job preparation. Here is a list of the data elements that were tested:

- On the job training (11% of the donor to recipient matches were decided randomly)
- Standing/walking (11%)
- Pushing with the upper body (14%)
- Peripheral vision (14%)
- Exposure to hazardous contaminants (14%)

Once a set of variances were calculated for both the current method and the proposed method, a few summary graphs were produced. To start, the following graph shows how many published variances were exactly the same regardless of the method. Slightly more than a third of the variances were exactly the same, while slightly fewer than half of the variances increased under the proposed method. Note that an increase in variances under the proposed method was the hypothesis (in other words, the current method underestimates the variances).

*Graph 1: Comparing the proposed method to the current method*

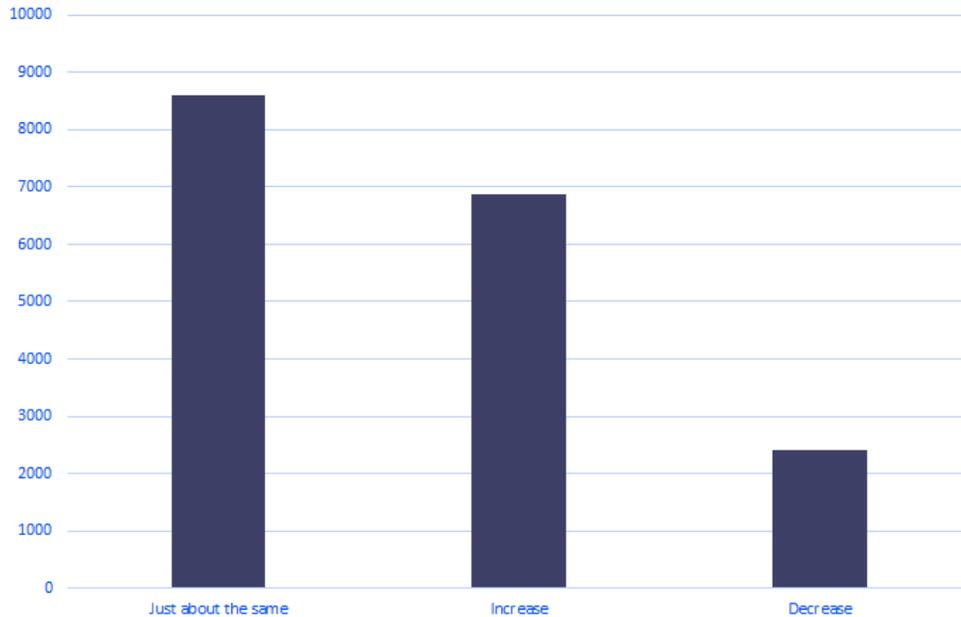
## How do the Proposed Standard Errors Compare to the Current Method?



Now, exact matching is great, but what if a very small amount of variation were allowed? Here is the same graph, only this time “no change” occurs at the 0.0001 level. Using this definition for “no change,” about half of the variances are “just about the same” while about 40% of the variances increase under the proposed method.

*Graph 2: Comparing the proposed method to the current method at the 0.0001 level*

## What About Comparing at the 0.0001 Level?



Now, to extend past summary-level information, a more nuanced comparison must be made. Suppose the following conditions for “almost the same” are considered:

1. Identify absolute standard error differences that are smaller than 0.5
2. Identify relative standard error differences that are smaller than 3%

Note that relative standard errors are defined as the absolute standard error difference divided by the full sample estimate value.

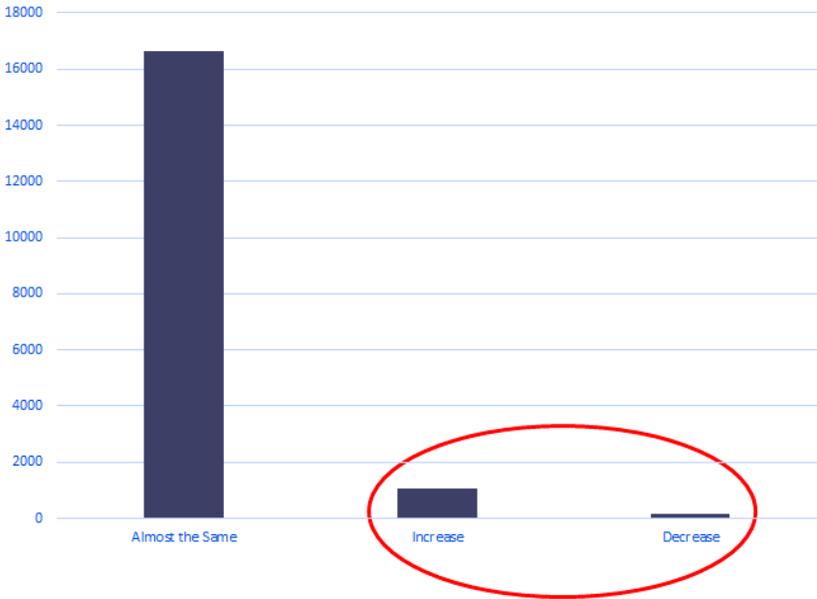
Both of these criteria are rather stringent individually. Here, the two criterion will be combined so that a standard error that is “almost the same” must meet both criteria.

Now, there are several known ways to test for equality of variances – an F Test, Levene’s Test, Brown-Forsythe’s Test, and Bartlett’s Test. These tests were not employed for this research because the ORS data does not acceptably satisfy the assumptions needed.

The following graph displays the results of this comparison, showing that now around 93% of the standard errors are “almost the same” as defined by the criteria above.

*Graph 3: A more nuanced comparison of the proposed method and the current method*

## How do the Proposed Standard Errors Compare to the Current Method?

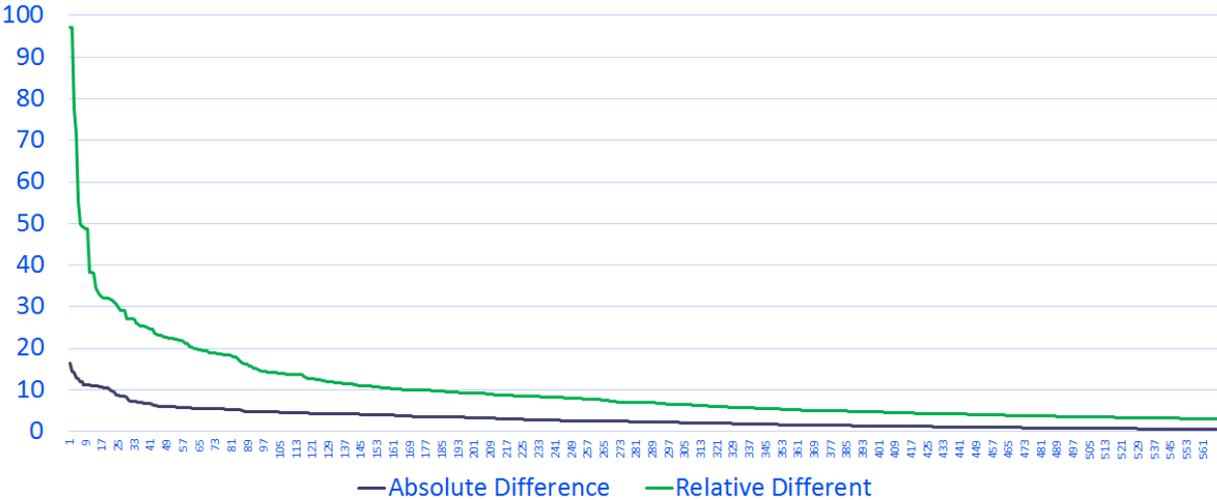


So, only 7% (circled above) of the standard error estimates have a difference deemed significant enough for further investigation. The following graphs and tables explore possible trends and explanations for why these standard errors stand out as different.

Below, Graph 4 shows the sorted absolute differences for percentage of worker estimates. Only a few on the left side of this graph are wildly different, while the rest of the standard error differences taper off fairly quickly.

*Graph 4: Absolute and relative differences for percentage of workers estimates*

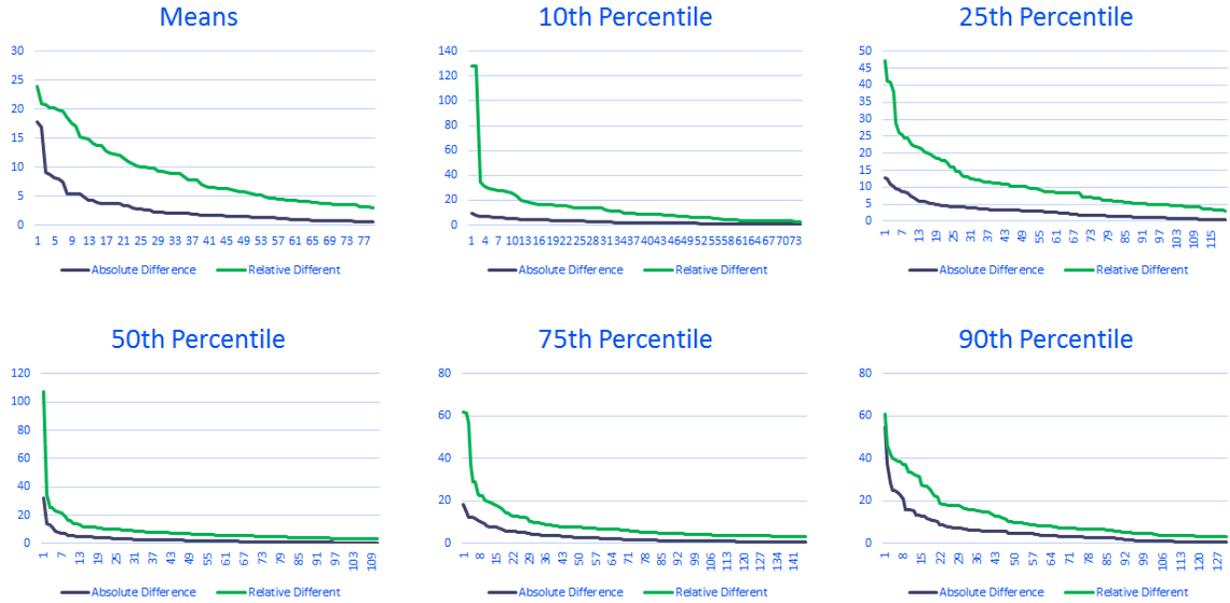
### Absolute vs Relative Differences for Percentage Estimates



Graph 5 is similar to Graph 4, though here the rest of the variance estimates have been graphed. Again, the same patterns tends to emerge – there are a few variance estimates with very large differences, while the rest of the differences are still fairly small.

The mean estimates seem like one exception, however a closer look at the scale shows that the largest relative differences are between 15% and 20%. These differences are significant, but relative to all other differences, they are not extremely significant.

Graph 5: Absolute and relative differences for other types of estimates



Are there trends among occupations? Table 1 below shows the top 14 occupations with the most prevalent (among the 7% of standard errors that were different) standard error differences. The “count” refers to the number of standard error differences and the “total” refers to the total number of published estimates associated with the occupation.

There does not seem to be much of an occupational trend, as the top 14 occupations are fairly varied and most of them do not have very many published estimates. Note that the maximum amount of estimates per occupation is 484.

Table 1: Occupations with the most standard error differences

SOC Label	Count	Total	Percent
Municipal Firefighters	30	94	32%
Computer Network Support Specialists	26	59	44%
Sales Representatives, Wholesale and Manufacturing, Technical and Scientific Products	25	70	36%
Sales Agents, Financial Services	24	62	39%
Child, Family, and School Social Workers	23	80	29%
Paralegals and Legal Assistants	22	60	37%
Tellers	22	78	28%
Office and Administrative Support Workers, All Other	22	62	35%
Registered Nurses	21	174	12%
Mental Health and Substance Abuse Social Workers	19	48	40%
Elementary School Teachers, Except Special Education	18	112	16%
Shipping, Receiving, and Traffic Clerks	18	79	23%
Stock Clerks- Stockroom, Warehouse, or Storage Yard	18	68	26%
Machinists	17	57	30%

Are there any trends by ORS data element? Table 2 shows the five data elements that were studied. Only pushing and pulling with the upper body stands out with 15% of the variance estimates being different enough to require further investigation.

*Table 2: Data elements and corresponding standard error differences*

Data Element	Count	Total	Percent
Standing/walking	531	7917	7%
Pushing/pulling with the upper body	245	1602	15%
Post-employment training	154	3773	4%
Peripheral vision	108	1173	9%
Hazardous contaminants	86	1445	6%

Are there any trends by estimate type? Table 3, below, shows the estimate types with the corresponding prevalence of standard errors that were different enough to require further investigation. The differences are evenly spread among the estimate types, and the highest prevalence of differences is just 8%. So there does not seem to be a trend among estimate types.

*Table 3: Estimate types and corresponding standard error differences*

Estimate Type	Count	Total	Percent
Percentage of Workers	568	7349	8%
75th Percentile	146	1726	8%
90th Percentile	130	1700	8%
25th Percentile	120	1682	7%
Median	111	1660	7%
Mean	79	2096	4%
10th Percentile	74	1656	4%

Are there any trends by data element type? Table 4, below, shows that there are no discernable trends by data element type. The highest prevalence of standard error differences by data element type is just 8%.

*Table 4: Data element types and corresponding standard error differences*

Data Element Type	Count	Total	Percent
Physical Demands	925	11092	8%
Job Preparation	217	5332	4%
Environmental Conditions	86	1445	6%

## 6. Summary of the Results

In conclusion, the research illustrates that the effect of ties on the variance estimates is minimal. Most, 93%, of the variance estimates did not substantially change as a result of using proposed method. In addition, there were no obvious trends in the 7% of estimated variances that were different enough to require further investigation. Trends were examined by occupation, specific data element, estimate type, and data element type. For the 7% that were different, the proposed method produced larger variances than the current method.

Future research could include an extension of scope (by analyzing more ORS data elements) and a more formal simulation study on the effect of breaking ties in a deterministic imputation procedure.

## References

- [1] See and Social Security Administration, Occupational Information System Project
- [2] See U.S. Department of Labor, Dictionary of Occupational Titles (DOT)
- [3] See Occupational Information Development Advisory Panel, 2010
- [4] See O\*Net Online, <http://www.onetonline.org/>
- [5] See Occupational Requirements Survey website, <http://www.bls.gov/ors/>
- [6] Quarterly Census of Employment and Wages, <http://www.bls.gov/cew/>.
- [7] North American Industry Classification System, <http://www.census.gov/eos/www/naics/>.
- [8] Standard Occupational Classification System, <http://www.bls.gov/soc/>.
- [9] Ferguson, Gwyn R., McNulty, Erin. 2015. Occupational Requirements Survey Sample Design. In JSM proceedings, Economic Data: CPI, PPI, NCS Section. Alexandria, VA: American Statistical Association.
- [10] Righter, Leland, Rhein, Bradley 2016. Imputation Methodology for the Occupational Requirements Survey. In JSM proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association.
- [11] 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.
- [12] 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.
- [13] Rhein, Bradley D., Ponikowski, Chester H. 2016. Variance Estimation for the Occupational Requirements Survey. In JSM proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association.
- [14] Andridge, Little. 2010. A Review of Hot Deck Imputation for Survey Non-response. International Statistical Review.

**Note: Any opinions expressed in this paper are those of the author(s) and do not constitute policy of the Bureau of Labor Statistics.**