Visual Validation of Estimates from the Occupation Requirements Survey November 2017

Lacoya Theus and Andrew Kato

U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington, DC 20212

Abstract:

The Bureau of Labor Statistics (BLS) has completed the second year of collection of the Occupational Requirements Survey (ORS), which provides information on the physical demands, mental requirements, education and training, and environmental conditions of an occupation. These elements, 77 in total, provide a robust set of estimates that offer a vivid description of occupational requirements in the U.S. economy. Since ORS is a newly established survey, it was necessary to develop a set of processes that would detect anomalous estimates. To validate these estimates a team created an interactive visualization tool using Tableau, to evaluate the estimates against our expectations and identify anomalies in the dataset. Estimates flagged as anomalous were isolated and investigated to determine if they met the criteria for suppression. This paper describes the validation procedures that estimates undergo to determine if the ORS estimates are fit for use, with emphasis on the visualization tools used in analyzing the estimates.

1. Introduction

The Bureau of Labor Statistics (BLS) along with the Social Security Administration (SSA) entered into an interagency agreement in 2012 to provide information that could assess the occupational needs of workers in the economy. This resulted in the development of the Occupational Requirements Survey [1], which seeks to provide updated information on demands of specific occupations. ORS is an establishment-based survey collected by BLS field economists across the country via in-person visits and other forms of direct contact with respondents (e.g. phone). During the collection period, regional staff collect 77 data elements characterizing jobs at respondents' workplaces divided into 4 categories: education and training, mental requirements, physical demands, and environmental conditions.

The specific format of information collected for each data element varies. The four categories are collected in the form of specific data points, ranges, and in yes/no values. Respondents previously had the ability to provide data for the amount of time required for a physical demand or exposed to an environmental condition by specifying a particular number of hours, percentage of the workday, or even a range of hours or percentages. The full range of elements collected and the forms used by BLS field economists are available online [2].

A complicating factor in evaluating ORS data is the mixed nature of the survey structure. While the sampling processes are grounded in establishment-based survey concepts using familiar stratification strategies (e.g. by industry, by geographical area, etc.), micro-data collected from respondents about jobs at those targeted workplaces are grouped for

estimation using an occupation-based concept. Estimates produced by the ORS program are for occupations as classified by the Standard Occupational Classification (SOC) system [3]. Therefore, data collected from a single respondent may contribute to many different estimation cells; seen from the standpoint of validating estimates, one estimation cell draws upon data collected from many different sampling cells. These complex inter-relationships between and across sampling cells, and estimation cells present a challenge to reviewers of the micro-data, and the validators of the estimates.

Consequently, BLS has explored data visualization platforms to focus review efforts and reduce the amount of staff resources necessary to deal with such a complicated dataset. This paper describes the path taken by BLS to implement visualization methods at the estimate validation stage. In this paper, we briefly introduce the difference between micro-data review and estimate validation, discuss some of the early difficulties encountered in ORS validation, and finally lay out our road map forward and some lessons learned.

2. Data Validation in ORS

Data quality is assessed often and in detail at many stages of the ORS survey cycle: several times at the micro-data level [4] [5] during collection, and again after collection is closed at an aggregate level (estimate validation). The purpose of validation is to analyze the dataset of estimates to ensure that survey processes and methods are working as intended. During this process we ensure that estimates are aligned with expectations and provide additional analysis to justify data that are outside of our expectation. While regular micro-data review evaluates individual survey responses (commonly referred to as quotes) with an aim of identifying and preventing estimation errors, validation evaluates estimates, quotes or processes that may need attention.

After the data is collected and reviewed micro-data are used to calculate weighted estimates, the output, a dataset of estimates and standard errors, are scanned to see if the results are consistent with expectations. Anything unusual that may have been difficult to spot among individual records stands out at this stage. Estimates that do not meet expectations are further investigated to either confirm the unusual result as Fit-for-Use (FFU) (e.g. estimates that are available for publication) or alert validators to an issue with collection, estimation procedures, or with other survey methods or processes.

This validation of estimates involves four major steps:

1. Set expectations - using data and research from studies and programs concerned with occupational requirements such as the US Department of Labor's Dictionary of Occupational Titles (DOT) [6], the aforementioned O*Net datasets, and occupational data from the BLS Occupational Employment Statistics program, ORS analysts can form rough approximations on what values are expected from ORS estimation.

2. Identifying anomalies - BLS analysts compare current estimates to the expectations for the estimates. While some estimates will be near expectations within some allowable tolerance, others may be unexpectedly much higher or

much lower than what was projected. For example, a validator may not be surprised by a high proportion of Nurse Practitioners being required to have post-graduate college degrees, but it would not be expected that a high proportion of Bakers are required to have a post-graduate college degree.

3. Investigation of anomalies – Once unusual estimates are identified, analysts examine the micro-data within the cell to understand why the calculated value differed from expectations. If the quotes in an examined cell appear to contain accurately coded data backed by information provide by field economist, and results from the review of the micro data is checked, the data are confirmed as valid.

4. Documentation – The outcomes for any estimates flagged in step 2 and examined in detail in step 3 are recorded in internal validation reports. These reports are then used to inform decisions on whether to possibly suppress questionable estimates if the errors as appropriate.

The most useful guide for setting ORS validation expectations in step 1 would be prior results from the same survey, but such data is unavailable at this time. As pointed out in the *Validation of Estimates in the Occupational Requirements Survey: Analysis of Approaches*, the ORS is a new survey without a long history of published estimates to draw on [7]. Due to differences in coding structures, scope of coverage, and collection methodology, information from outside sources such as the DOT or O*Net are helpful but ultimately limited in applicability for assessing how Fit-For-Use any particular ORS estimate may be. Until additional data accumulates over time to drive adaptive expectations for computational methods, resource-intensive data analyzing by staff assigned to perform estimate validation is the only option.

3. Evolutionary approach to visualization for ORS

The full output dataset of estimates for the ORS requires substantial staff time to review manually, even with a large staff of analysts, due to its enormous size. For every basic cell designated by either a detailed occupation code or a range of SOC codes, all 77 data elements could potentially be estimated. For a particular data element, several characteristics or alternative categories may be estimated. Each of those specific SOC-element-characteristic combinations then possesses several attributes. For example, consider the physical demands requirement of reaching at shoulder height:

- First, ORS asks what percent of the occupation's workday is spent doing such reaching, and the weighted shares within a SOC cell in each of five ranges are computed as separate estimates: Not Present, Seldom (0 to 2 percent of the work shift), Occasional (2 to 33 percent), Frequent (33 to 66 percent), and Constant (above 66 percent).
- Conditional on reaching at shoulder height being an occupational requirement, the ORS asks if the position requires such reaching with one hand (an estimate) or both hands (also an estimate).

Thus, for this one data element of reaching at shoulder height, there are 13 potential estimates per SOC cell. In the first official ORS dataset published by BLS in December

2016, more than 180 SOC cells met minimum quality requirements to calculate estimates. Since each discrete estimate has an estimate value, a standard error, and a publishability status, there could have been more than 7,000 pieces of information to consider when reviewing just the shoulder reaching element. Spread across all 77 data elements and their linked characteristics, the volume of information to process rises rapidly into the hundreds of thousands of estimates, standard errors, and publishability status flags. Such an immense dataset is large and complex.

BLS adopted data visualization as a way to focus and accelerate estimate validation in ORS. At its core, visualization is about rearranging and organizing data in more useful ways to facilitate decisions. It is not a way to automate or replace the analyst, but involves presenting information in a manner that allows the analyst to sift through large volumes of information to find what they need to draw conclusions in less time. [8] It was important in ORS validation to be able to scan rapidly across two dimensions: the same variable for many different SOC codes, and many different variables for a single SOC code. The preliminary form this visualization took was similar to the table shown in Figure 1 below, which uses dummy data and publishability flags for a fictional 707 survey cycle.

Figure 1: Example Validation Table of the type used in 2016

	A	B	C	D	P	Q	R	S	T	U	V	W	X	Y
1	BASIC_CELL_ID	SOC_GROUP_LABEL	LOWER_SOC_CODE	UPPER_SOC_CODE	No_Ed_tot	No_Ed_litY	No_Ed_litN	HS_only	Assoc_deg	Bac_deg	Mast_deg	Prof_deg	Doc_deg	Oth_deg
8	1-0000-376-000	Elementary School Teachers, Except Special Education	25202100	25202100	.	.	.	15.1	0.6	51.6	2	.	.	0.4
9	1-0000-376-000	Elementary School Teachers, Except Special Education	25202100	25202100	.	.	.	P	P	P	F	.	.	P
10	1-0000-377-000	Middle School Teachers, Except Special and Career/Technical Education	25202200	25202200	.	.	.	7.7	1.9	56.6	5	.	.	.
11	1-0000-377-000	Middle School Teachers, Except Special and Career/Technical Education	25202200	25202200	.	.	.	F	F	P	P	.	.	.
12	1-0000-379-000	Secondary School Teachers, Except Special and Career/Technical Education	25203100	25203100	.	.	.	4.8	.	93	1.9	.	.	.
13	1-0000-379-000	Secondary School Teachers, Except Special and Career/Technical Education	25203100	25203100	.	.	.	F	.	F	P	.	.	.
14	1-0000-400-000	Teacher Assistants	25904100	25904100	2.3	2.5	.	51.1	18.5	1.8	.	.	.	0.5
15	1-0000-400-000	Teacher Assistants	25904100	25904100	F	P	.	F	P	P	.	.	.	P
16	1-0000-498-000	Registered Nurses	29114100	29114100	.	.	.	7.2	24.9	68.7	0.9	0.2	.	.
17	1-0000-498-000	Registered Nurses	29114100	29114100	.	.	.	P	F	P	P	F	.	.
18	1-0000-530-000	Licensed Practical and Licensed Vocational Nurses	29206100	29206100	1.5	1.5	.	50.1	30.4	2.1	.	.	.	1.5
19	1-0000-530-000	Licensed Practical and Licensed Vocational Nurses	29206100	29206100	F	F	.	F	P	F	.	.	.	F

Fake 707 Example Data

This early table format for validation is a standard table which looks and feels familiar to BLS analysts, allowing them to immediately use the tool without any special training. Instead of one SOC-element-characteristic per row, the data is rearranged to show one occupation cell per row with grouped characteristics for each element in the columns. The table includes the publishability status of each estimate immediately below the estimate value itself. This layout makes it possible to quickly view many rows vertically along a column to compare the experience of many occupations for the same element. It is also easy to scan horizontally across a single row to check all linked characteristics of

an element for a particular occupation or even multiple data element groups for that SOC code.

Over the course of validation activities in 2016, analysts used numerous visual elements to augment this table format including the use of color to highlight cell backgrounds or change font color, typeface changes such as italicization or boldface, and border selections around cells. These visual elements, which draw the attention of the viewer, transform the table from a plain text table into something more useful, featuring targeted information. Such tables that begin with normal text and then overlay visual cues and elements are in fact considered a staple type of data visualization called highlight tables [9].

The table format in this application did not meet the particular needs for ORS. Thus, BLS worked on migrating the highlight table concepts from that early version into a full-featured visual validation tool in Tableau, a software package specifically designed for data visualization. Figure 2 shows an early prototype of such a migrated table, using the same dummy dataset as before for fictional survey cycles numbered, Control Group 706 and 707.

Figure 2: Early prototype of Tableau highlight table for ORS validation

			Minimum education level						
Basic Cell Id	Control Gr..	Soc Group Label	NONE	HIGH SCHOOL	ASSOCIATE'S	MASTER'S	PROFESSION..	DOCTORATE	OTHER
1-0000-376-000	706	Elementary School Teachers, Except Special Ed..		11.20	0.70	3.10			0.30
	707	Elementary School Teachers, Except Special Ed..		15.10	0.60	2.00			0.40
1-0000-377-000	706	Middle School Teachers, Except Special and Car..		9.80	1.30	3.80			
	707	Middle School Teachers, Except Special and Car..		7.70	1.90	5.00			
1-0000-379-000	706	Secondary School Teachers, Except Special and..		4.30		3.60			
	707	Secondary School Teachers, Except Special and..		4.80		1.90			
1-0000-400-000	706	Teacher Assistants	2.70	99.10	16.10				0.40
	707	Teacher Assistants	2.30	51.10	18.50				0.50
1-0000-498-000	706	Registered Nurses		5.30	24.30	1.00	0.20		
	707	Registered Nurses		7.20	24.90	0.90	0.20		
1-0000-530-000	706	Licensed Practical and Licensed Vocational Nurs..	1.30	37.90	31.50				1.20
	707	Licensed Practical and Licensed Vocational Nurs..	1.50	50.10	30.40				1.50
1-0000-548-000	706	Nursing Assistants	27.40	60.80	0.80				0.20
	707	Nursing Assistants	38.30	82.30	0.50				0.10
1-0000-556-000	706	Medical Assistants	0.40	87.20	15.20				0.80
	707	Medical Assistants	0.40	90.30	15.90				1.20

Better control of visual elements in a software package (Tableau) suited to creating data visualizations increases the table's value to validators. Here, the highlight in orange (pass) and blue (fail) are being used to denote publishability status without having to explicitly declare the value in text. Use of visual elements makes it possible to fit more information into a less cluttered view; note that this table shows the value for two hypothetical years of data, for 707 and its predecessor 706 sample. Such a visualization demonstrates an effortless way for a highlight table to show prior year information, providing context to help frame step 1 expectations for validators while doing step 2 identification of anomalous estimates. Now that the ORS has ended its second survey cycle, validators are able to review the rates of change between 2016 and 2017 estimates.

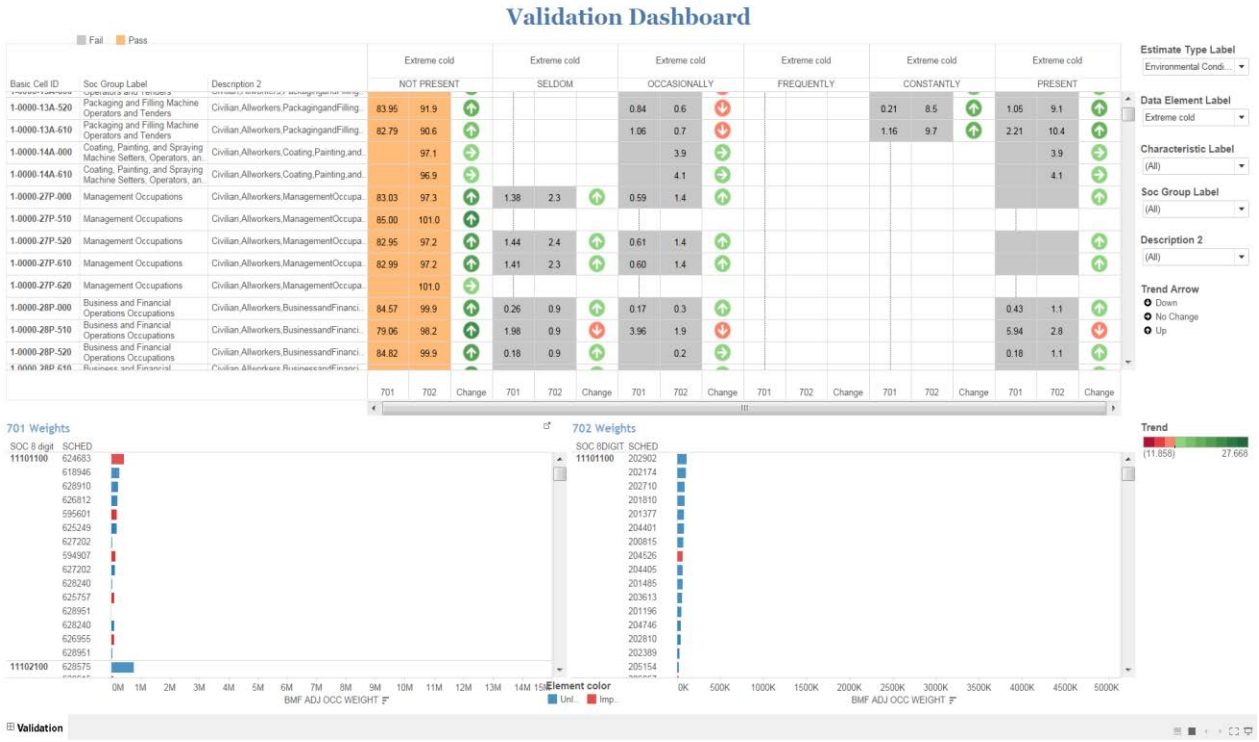
Now that the core highlight table view satisfies the first requirement, current work seeks to provide users with greater control of the dataset. User-selectable filters which could

flag cells with icons or other visual cues whenever they satisfy pre-set conditions such as "publishable and has increased or decreased more than ten percent from the prior year estimate" are in development. The exact criteria used in such filters will be informed by the experience of BLS validation analysts as they learn each year to spot anomalous estimates. Such learning can be embodied in the tool as a filter to pre-identify such cells as very likely to warrant investigation in step 3.

Ongoing efforts to improve and refine the data visualizations used in ORS validation also led to development of a Tableau dashboard that links the highlight table view of the dataset to other panels on the dashboard, which displays information by occupation for cells clicked by the user. Providing such details-on-demand extends the functionality of the visualization tool to step 3 of the validation process. Opening and reviewing the weights, imputation status, data element values, and other attributes of individual quotes for step 3 in an integrated fashion on the same dashboard view as the step 2 highlight table adds savings by eliminating the need to switch to other applications to review micro-data.

The dashboard displayed below is an illustration of what validators use during the current cycle of estimate validation. The left hand rows displays the SOC-code, which provides the validators with the occupational title information. The columns display the variable titles and the rate of change that an estimates experienced from one year to the next. The rate of change is displayed both by the arrows and the color of arrow. The up green arrows indicates an increase of the estimate from last year's collection cycle; while a green sideways arrows indicates no change, and a red downward arrow indicates a decrease. On the right side of the dashboard, validators have access to toggle the estimates based on the estimates type, data element label, SOC-group, and the description name. The bottom of the dashboard displays the weights, the length of the bars indicate how heavily weighted the data are and the color indicates if there are an imputed data points.

Figure 3: The Current Version of the Validation Dashboard



The final step required for the validation team to document our results within the validation tool is being investigated but reserved for future iterations of the tool. Once the "overview first, zoom and filter, then details-on-demand" [10] aspects are completed to the users desires, the ability to add notes and export work lists from the tool would be essential. Tableau is intended for use as a read-only viewer of data, though, so data creation in the form of analyst comments may be tricky to add.

4. Conclusion

The intention behind the creation of the visual interactive dashboards was to provide users with information to validate estimates seamlessly with review at the micro-data level. This process has gone through many iterations before its current version. We began with a rudimentary understanding of data visualizations and the goal of the tool was to simply provide the user with an overall understanding of the estimates and the relationships that existed between elements. Looking towards the future it is our expectation that the new technology will offer new insights into the data and allow a full integration between ORS review and validation. Estimate validation in BLS' Occupational Requirements Survey provides a great opportunity to use visualization techniques to obtain higher quality validation of estimates at lower time cost by enabling analysts to focus their efforts better. The evolutionary approach taken by the ORS program in implementing visualization allowed analysts to adapt to using visualization methods in an otherwise familiar context the first year, paving the way for transition to a

fuller implementation of those techniques in a more specialized platform. BLS expects to continue augmenting the Tableau tool for validation of future ORS survey cycles.

References

[1] More detailed explanations about what the survey entails as well as its purpose can be found on the ORS website https://www.bls.gov/ors/

[2] The collection forms and collection procedures manual used by the ORS program are available at: https://www.bls.gov/ncs/ors/orspub.htm

[3] The ORS uses occupation codes from the 2010 Standard Occupational Classification system supplemented by additional detail established by O*Net. Official SOC codes carry six digits of detail, while O*Net occupation codes add a further layer by extending the code length to eight digits. Detailed explanation of the official six-digit SOC 2010 codes can be found on the BLS web page at: https://www.bls.gov/soc/For more information on the O*Net-SOC 2010 Taxonomy, see O*Net's documentation at: https://www.onetcenter.org/reports/Taxonomy2010.html

[4] Meharenna, Ruth. 2015. Occupational Requirements Survey (ORS) Data Review Process in *JSM Proceedings*, Government Statistics Section. Alexandria, VA: American Statistical Association. Available at: https://www.bls.gov/osmr/abstract/st/st150040.htm

[5] Brown, Karen and Harney, Tamara. 2015. Building Quality Assurance for the Occupational Requirements Survey in JSM Proceedings, Quality and Productivity Section. Alexandria, VA: American Statistical Association. Available at: https://www.bls.gov/osmr/abstract/st/st150030.htm

[6] U.S. Department of Labor, Employment and Training Administration (1991), Dictionary of Occupational Titles, Fourth Edition, Revised 1991

[7] Smyth, Kristin. 2015. Validation of Estimates in the Occupational Requirements Survey: Analysis of Approaches in JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association. Available at: https://www.bls.gov/ncs/ors/validation.pdf

[8] See for example, David McCandless' TEDGlobal 2010 talk, where trying to deal with mass quantities of data is overwhelming "if you look at it directly, it's just a lot of numbers and disconnected facts. But if you start working with it and playing with it in a certain way, interesting things can appear and different patterns can be revealed." Available at: https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization/transcript?language=en#t-56000

[9] See for example, *Tableau Essentials: Chart Types – Highlight Table* by Carly Capitula, available at: https://www.interworks.com/blog/ccapitula/2014/08/06/tableau-essentials-chart-types-highlight-table

[10] Shneiderman, Ben. 1997. A Grander Goal: A Thousand-Fold Increase in Human Capabilities in Educom Review, 32, 6 (Nov/Dec 1997), 4-10. Available at: http://www.ifp.illinois.edu/nabhcs/abstracts/shneiderman.html