An Optimization Approach to Reconciling Sample Allocations November 2017

David Piccone[1] and Matthew Dey[1]
The U.S. Bureau of Labor Statistics, Washington D.C.

Abstract
As part of a larger project, a research team at the Bureau of Labor Statistics (BLS) created an alternative sample design for the Occupational Employment Statistics (OES) survey. There are three sample allocations for the new sample design, each geared towards improving the estimator in different ways. There is an efficient allocation that aims to lower the sampling error of the OES estimates, and two minimum allocations that set a lower sample size threshold for area and industry domains. Each of the three sample allocations are stratified designs, however they use different strata definitions. This paper describes how we reconcile the three allocations using an optimization approach.

Key Words: optimization, stratified sample allocation, establishment survey, power Neyman allocation, minimum allocation, occupational employment statistics

1. Background

The Occupational Employment Statistics (OES) survey collects occupational employment and wage information from a sample of business establishments found in the 50 United States, the District of Columbia, Guam, Puerto Rico and the Virgin Islands. These data items are used to create point-in-time estimates of occupational employment levels and hourly and annual mean wage estimates for over 800 detailed Standard Occupational Classification (SOC) occupations. These estimates are calculated nationally, by state and territory, by detailed Metropolitan Statistical Areas (MSAs) and non-MSA areas called Balance of State (BOS) areas, and by North American Industry Classification System (NAICS) industries. The OES program publishes estimates annually.

The current estimation methods for the OES program require a very large sample size in order to produce detailed area and industry occupational estimates. The sample is selected and collected based on a three year survey cycle, where approximately 400,000 establishments are sampled annually. To the extent possible, the establishments selected in any given year are excluded from selection in the next two preceding years. The OES sample is selected using a probability proportional to size (PPS) sampling scheme, where an establishment's employment is its measure of size. To provide adequate geographical, industrial, and occupational coverage, OES combines three years of sample to produce estimates. This is done by using a rolling three year cycle, where a current annual sample is rolled in to replace an older annual sample selected three years prior. Approximately 1.2 million sampled establishments are used for any given set of estimates (BLS Handbook of Methods, 2017).

A major drawback of using this type of sample design is that it prevents the OES from publishing time-series estimates. The reason being that establishments are sampled once every three years, causing even the most influential establishments to provide data only once in any three year period. In the intermediate years the data from these establishments are updated using certain assumptions. The yearly changes that occur within these establishments are very important for change estimates, and without capturing these changes every year, OES time series estimates will suffer from biases. A research team at

[1] Any opinions expressed in this paper are those of the author(s) and do not constitute policy of the Bureau of Labor Statistics.

the Bureau of Labor Statistics (BLS) has worked on redesigning the OES survey's sample design and estimation methods in order to produce valid time-series estimates.

The focus of this paper will be on an offshoot project that arose while researching the new OES sample design. To improve the time-series estimation methods, we developed three separate stratified sample allocations. The first allocation is geared towards selecting an efficient sample, where strata with more employees and larger occupational heterogeneity are allocated more sample. The other two allocations focus on ensuring that a minimum amount of sample is selected in every NAICS industry and MSA-BOS area. This paper will explain how we tested several different ways of reconciling these allocations. In the remaining sections we will discuss the current allocation methods, our proposed allocation methods, an optimization approach to reconciling sample allocations, the results from testing our proposed methods, and a conclusion.

2. Current Sample Allocation

The current OES sample design is similar to the proposed sample design in that it uses an efficient allocation and a minimum allocation. Both allocations stratify by state, MSA-BOS area and four-digit NAICS industry (NAICS4). The efficient allocation uses the Power Neyman allocation:

$$n_h = n\frac{\sqrt{X_h}\,S_h}{\sum_{all\ h}\left(\sqrt{X_h}\,S_h\right)} \tag{2.1}$$

Where,

n_h = the amount of sample allocated to stratum h (State by MSA-BOS areas by NAICS4)
n = the national sample size
X_h = the number of employees in stratum h
S_h = the measure of occupational employment variability within stratum h

The Power Neyman allocation provides larger sample sizes to strata that have more employees and more occupational heterogeneity. This aims to drive down the sampling variance of the OES estimates. The Power Neyman allocation is similar to the Neyman allocation, except the measure of size (employees) is raised to the power of ½. By using the Power Neyman allocation over the Neyman allocation, sample is shifted from the largest strata to the mid-sized and small strata, which allows for more precise estimates for the smaller domains at the expense of some precision in the largest domains. The minimum allocation uses the following rules:

$$m_h = \begin{cases} N_h & if\ N_h \le 3 \\ 3 & if\ 4 \le N_h \le 11 \\ 6 & if\ N_h \ge 12 \end{cases} \tag{2.2}$$

Where,

m_h = the amount of sample allocated to stratum h
N_h = the number of frame units in stratum h

The minimum allocation aims to help estimates meet the OES confidentially criteria in order to increase the total number of published OES estimates.

To reconcile the two allocations, the final sample allocated to each stratum is set to the maximum of the Power Neyman and minimum allocations. After the initial reconciliation, the overall sample size is larger than the target sample size. An iterative process is used to systematically adjust the national sample size value (n) used in formula 2.1 until the overall reconciled sample allocation is close enough to the target sample size. The current way of taking the maximum sample then iterating to reconcile the sample allocations will be referred to as the "simple approach" for the remainder of the paper.

3. New Sample Allocation

The new sample design uses an efficient allocation and two minimum allocations. Each of the three allocations uses a different stratification plan. The efficient allocation stratifies by State, aggregate area and NAICS4 industry. Aggregate areas are combinations of similar MSA-BOS areas based on how close the areas are geographically to each other within a given state. The largest MSA-BOS areas are not aggregated with any other areas. The proposed efficient allocation uses the Neyman allocation:

$$n_k = n \frac{X_k S_k}{\sum_{all\ k}(X_k S_k)} \qquad (3.1)$$

Where,

n_k = the amount of sample allocated to stratum k (State by Aggregate area by NAICS4)
n = the national sample size
X_k = the number of employees in stratum k
S_k = the measure of occupational employment variability within stratum k

There are two separate minimum allocations: 1.) the industry minimum allocation and 2.) the area minimum allocation. The goals of these allocations are different.

The industry minimum allocation aims to ensure we collect at least three observations for the most common occupations within each detailed 6-digit NAICS industry. The common occupations are the ones that make up the top 90th percentile of employment within the industry. We used previously collected OES micro data to determine which occupations are found in different size classes (i.e. groups of similarly sized establishments) within each industry. Since the OES is selected using a PPS sample, we could determine the expected number of sample units that would fall in each size class within each industry, given some sample size. By knowing the occupations found in each size class and the expected percentage of sample units that would fall in each size class, we could determine the likelihood of collecting each of the common occupations, given a sample of only one establishment. These likelihood measures would be greater than zero but less than or equal to one. The inverse of this likelihood measure is the expected sample size required to collect one observation of the common occupations. We multiplied this value by three to get the expected sample size needed to collect three observations for each common occupation. The final industry minimum allocation is the maximum value of these expected sample sizes within each 6-digit NAICS industry.

The area minimum allocation aims to increase the sample size for areas where there are large area effects on occupational employment levels. The two main predictors for occupational employment levels at an establishment are industry and establishment size. After controlling for these variables, area usually plays a small role in predicting occupational employment. However, there are some areas where there is a larger than

normal area effect. We allocate the area minimum allocation to detailed MSA-BOS areas proportional to the area effect for each MSA-BOS. The overall sample size for the area minimum allocation is set to equal the overall sample size of the industry minimum allocation.

4. New Approach to Reconciling Allocations

Unlike the current sample design, the proposed design has three allocations that are each using a different stratification plan. Figure 4.1 shows a visual representation of the three different stratification plans.

Figure 4.1: Visual Representation of the three different stratification plans for the proposed sample allocation

		MSA/BOS Detailed Areas									
		11260	21820	200001	200002	200009	38060	…	7800001	7800009	6-dig Industry Mins
6-Digit NAICS	113310							…			
	115111							…			
	115112							…			
	115113							…			
	115114							…			
	115115							…			
	115116							…			
	115210							…			
	…	…	…	…	…	…	…	…	…	…	…
	999200							…			
	999300							…			
	Area Mins							…			

The different shaped rectangles within the matrix in Figure 4.1 represent the stratification plan for the efficient allocation. The different widths of the rectangles represent the MSA-BOS aggregation that occurs to create the aggregate areas, and the different heights of the rectangles represent the aggregation that occurs for 4-digit NAICS industries. The right-most rectangles represent the industry minimum allocation strata, where each box represents a 6-digit NAICS industry. The bottom-most rectangles represent the area minimum allocation strata, where each box represents an MSA-BOS area.

The first step for reconciliation is to summarize the efficient allocation to the State by MSA-BOS area by 6-digit NAICS detail level so that we can easily aggregate to the detailed MSA-BOS areas and 6-digit NAICS industries. We will refer to this as the "detailed strata" level. Since the OES selects a PPS sample it is straightforward to distribute the efficient allocation to the detailed strata level. We first calculate a sampling interval (SI) for each efficient allocation stratum by dividing the measure of size (employment level) by the number of sample units allocated. Then we find the measure of size for each detailed stratum within the efficient allocation stratum. The sample size for each detailed stratum is its measure of size divided by the sampling interval. Figure 4.2 shows an example of how to break-out the efficient sample allocation within a particular efficient allocation stratum:

Figure 4.2: An example of how to break-out the efficient sample allocation to the detail strata level

		Aggr Area = 78	
		7800001	7800009
Aggr Ind = 8139	813910	X = 100	
	813920	n = 10	
	813930	SI = 10	

+

		Aggr Area = 78	
		7800001	7800009
Aggr Ind = 8139	813910	X = 20	X = 15
	813920	X = 30	X = 20
	813930	X = 10	X = 5

→

		Aggr Area = 78	
		7800001	7800009
Aggr Ind = 8139	813910	n = 2	n = 1.5
	813920	n = 3	n = 2
	813930	n = 1	n = 0.5

NOTE: X = measure of size, n = sample allocated, and SI = sampling interval: X/n

Once the efficient allocation is summarized at the detailed stratum level, it becomes easy to set up the reconciliation as an optimization problem. The goal of the reconciliation is to preserve the efficient allocation as much as possible while meeting the area and industry minimums. The minimums are met by increasing or decreasing the efficient allocation sample size in each detailed stratum by using adjustment factors. For our optimization problem we want to find the optimal set of adjustment factors in order to minimize an objective function that measures the distance between the efficient allocation and the reconciled allocation. The solution to the optimization problem is constrained by the following rules: 1.) the overall number of sample units allocated in the reconciled and efficient allocations must be equal, 2.) each adjustment factor cannot cause the reconciled allocation sample size to be less than or equal to 0 or greater than the number of frame units available to select, 3.) the reconciled allocation aggregated to MSA-BOS areas must be greater than or equal to the area minimum allocation values, and 4.) the reconciled allocation aggregated to 6-digit NAICS industries must be greater than or equal to the industry minimum allocation values. The notation for the optimization problem is as follows:

n_{ij}^{Eff} = efficient allocation sample size for cell ij, defined by 6-digit NAICS industry i and MSA-BOS area j

α_{ij} = adjustment factor for cell ij

$n_{ij}^{Opt} = n_{ij}^{Eff} \times \alpha_{ij}$ = Optimally reconciled allocation sample size for cell ij

N_{ij} = frame units in cell ij

M_i^{NAICS} = minimum sample size for 6-digit NAICS industry i

M_j^{MSA} = minimum sample size for MSA-BOS area j

Using this notation, the optimization problem can be summarized as minimizing objective function $f(\boldsymbol{\alpha})$ constrained by:

$$\sum_i \sum_j n_{ij}^{Eff} - \sum_i \sum_j \left(n_{ij}^{Eff} \times \alpha_{ij} \right) = 0 \qquad (4.1)$$

$$\alpha_{ij} \leq N_{ij} / n_{ij}^{Eff} \qquad (4.2)$$

$$\sum_j \left(n_{ij}^{Eff} \times \alpha_{ij} \right) - M_i^{NAICS} \geq 0 \quad \text{for } i = 1,2,\dots,I \qquad (4.3)$$

$$\sum_i \left(n_{ij}^{Eff} \times \alpha_{ij} \right) - M_j^{MSA} \geq 0 \quad \text{for } j = 1,2,\dots,J \qquad (4.4)$$

Where,

$I =$ the total number of 6-digit NAICS industries

$J =$ the total number of MSA-BOS area

5. Testing and Results

There are many different objective functions that we could use for the optimization approach. We tested six different objective functions, each providing a different distance measure between the final reconciled allocation and the efficient allocation. We compared the optimally reconciled allocations to two other allocations: 1.) the efficient allocation with no minimum allocations and 2.) the efficient allocation reconciled with the minimum allocations using the simple approach. By comparing the allocations with minimums to the efficient allocation with no minimums, we are able to measure the effect of the minimum allocations. By comparing the optimally reconciled allocations to the simply reconciled allocation, we are able to see if there are any gains by using the optimization approach over the current method.

To test the many different allocations, we made use of a simulated population that we created for the OES time-series project. The simulated population has occupational employment and wage data for all employees in every establishment in 18 states. The 18 states were chosen so that there was a mix of small, medium and large states spread across the different regions of the United States. We made use of industry and employment information on the OES frame to create a model that imputed occupational employment and wage information for every unseen (i.e. non-sampled and/or non-responding) establishment in the 18 states of our population. For establishments that provided data to OES within the 18 states, we used their occupational data as-is. We created the simulated population for five different time periods: 2005 to 2009. For the research presented in this paper we only used the 2007, 2008, and 2009 simulated population.

The simulated population gives us a measure of truth that we could use to evaluate the different allocations. For testing, we used a repeated sampling simulation study to measure how well each allocation does at estimating occupational employment. For each of the eight different allocations we selected 100 samples from the simulated population using Poisson sampling. We did this for 2007, 2008, and 2009 because the proposed estimation method uses a model that relies on three years' worth of OES survey data.

We first looked at how well the sampling procedures worked for each allocation. To do this, we used sampling weights to calculate weighted employment for each establishment found in our sample. We then aggregated this up overall and compared it to the overall population employment. Since we use Poisson sampling, the overall sample size, n, for each sample selected is a random variable that can range from 0 to N (the number of establishments in the population) (Hajek, 1958). The mean and variance of the sample size are:

$$E[\hat{n}] = \sum_{i=1}^{N} \pi_i \qquad (5.1)$$

$$V[\hat{n}] = \sum_{i=1}^{N} \pi_i(1 - \pi_i) \qquad (5.2)$$

Where,

$\hat{n} =$ the sample size using Poisson sampling

$\pi_i =$ the selection probability of establishment i in the population

A result of having a random variable sample size is that the weighted sample employment is not guaranteed to equal the population employment. However, it should be close. In Table 5.1 we show the average difference between the 100 weighted samples and the population for 2008.

Table 5.1: Weighted Sample vs. Population Employment Differences by Allocation Type

Allocation Description	Objective Fuction	Avg Diff	Avg Rel-Diff	% Samples with Neg Diff
Efficient Allocation with no mins	None	4,862	0.01%	44%
Simple Reconcilation	None	25,669	0.06%	42%
Optimal Reconcilation 1	$\sum_i \sum_j \left(1-\alpha_{ij}\right)^2$	-6,497,671	-15.94%	100%
Optimal Reconcilation 2	$\sum_i \sum_j \left(n_{ij}^{Eff}\left(1-\alpha_{ij}\right)\right)^2$	-5,700,449	-13.99%	100%
Optimal Reconcilation 3	$\sum_i \sum_j n_{ij}^{Eff}\left(1-\alpha_{ij}\right)^2$	6,702	0.02%	45%
Optimal Reconcilation 4	$\sum_i \sum_j n_{ij}^{Eff} ln\left(\frac{1}{\alpha_{ij}}\right)$	-6,573	-0.02%	55%
Optimal Reconcilation 5	$\sum_i \sum_j n_{ij}^{Eff} \frac{\left(1-\alpha_{ij}\right)^2}{\alpha_{ij}}$	27,152	0.07%	43%
Optimal Reconcilation 6	$\sum_i \sum_j \frac{\left(1-\alpha_{ij}\right)^2}{\alpha_{ij}}$	-20,297	-0.05%	56%

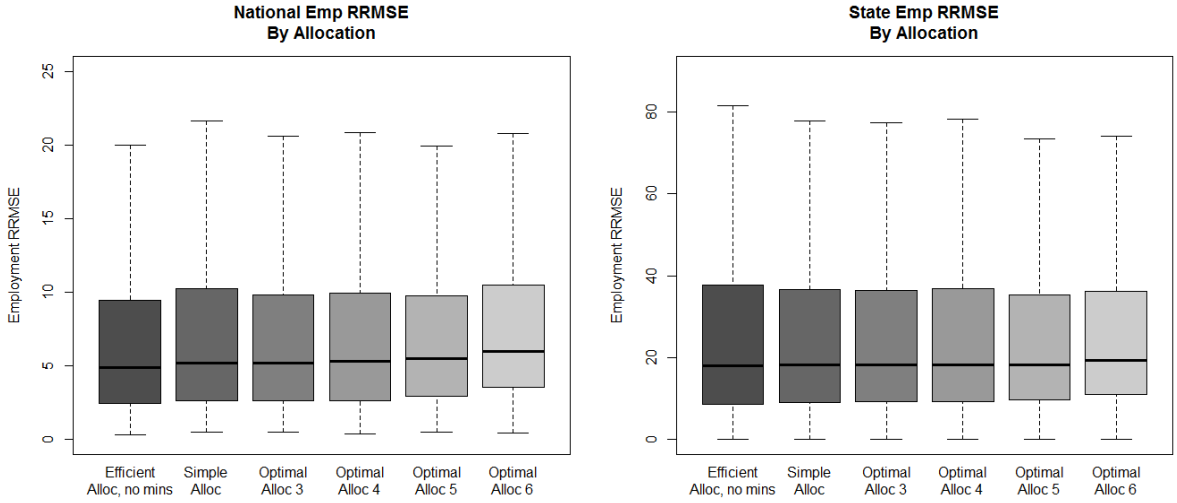
NOTE: The overall population employment is 40,750,723.

Ideally, an allocation would have an average relative difference close to zero and have about 50% of the samples with a negative difference. We found that the first two optimally reconciled allocations had large negative differences occurring in all 100 samples. The objective functions that we used for the first two optimally reconciled allocations adjusted the sample size for some detailed strata down to nearly zero, causing the establishments within these strata to have very small selection probabilities. This resulted in no sample being selected in many detailed strata, and therefore caused holes in the sample leaving large parts of the population unrepresented. It should be noted that in the unlikely event that an establishment with a very small selection probability is selected, it will have an extremely large sampling weight. Between the holes in the sample and the potential for very large sampling weights, the first two optimally reconciled allocations were dropped from consideration for the OES sample design.

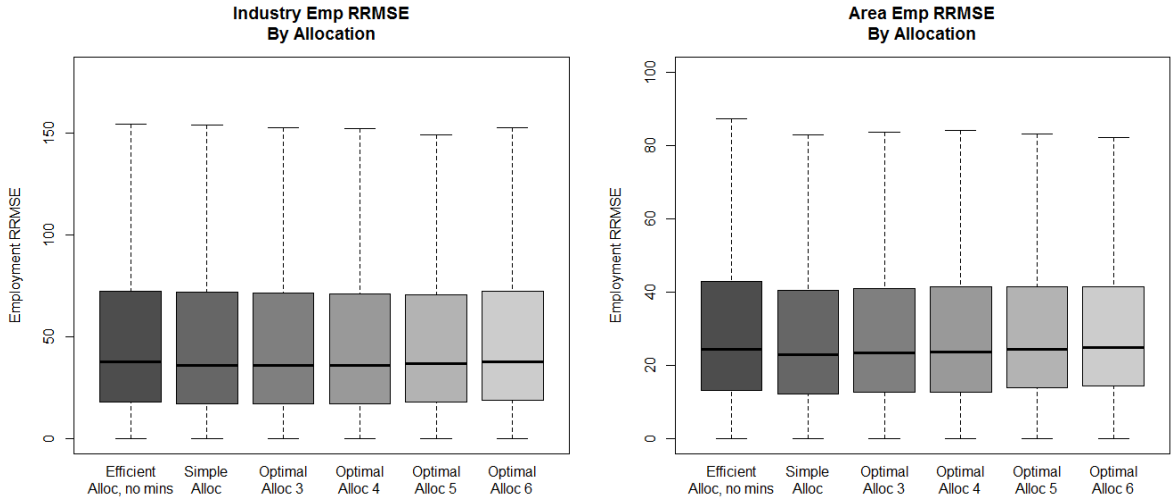
Next, we used the estimation methods proposed for the OES time-series project to produce 100 sets of occupational employment estimates for each allocation. We calculated these estimates for the four main OES estimation domains: 1.) National (i.e. all 18 states included in the simulation population), 2.) state, 3.) MSA-BOS area, and 4.) NAICS4 industry. Since we know truth from the simulated population, we were able to calculate

relative root mean squared error (RRMSE) statistics to measure the bias and variance of the employment estimates.

Graph 5.1: National and State RRMSE box plots by Allocation



Graph 5.2: Area and Industry RRMSE box plots by Allocation



Graphs 5.1 and 5.2 show that the RRMSE measures are very similar across the allocations within each estimation domain. The tables do show a very slight effect when adding minimums into the OES allocation. For national estimates, there is a small cost imposed when adding minimums, since the efficient allocation with minimums seems to be performing the best. For state and area estimates, there appears to be a small gain when adding minimums, since the efficient allocation appears to be performing the worst. For the industry estimates, there is no discernible effect of adding minimums.

For the national estimates, there are small improvements when using the optimal reconciliation approach for optimal allocations 3, 4 and 5 versus the simple approach. For the three sub-national domains, there does not seem to be gains from using the optimal reconciliation approach.

In addition to comparing the different estimates' RRMSE distributions for each allocation, we also looked at which allocation had the best RRMSE measure for the

individual estimates. In table 5.2 below, we show the percentage of times each allocation had the smallest, or tied with the smallest, RRMSE measure. These percentages are broken out by the size of the estimate. Small estimates are occupational employment estimates in the bottom 25th percentile, medium estimates are estimates between the 25th and 75th percentiles, and large estimates are estimates above the 75th percentile.

Table 5.2: Percent of Times each Allocation has the smallest RRMSE measure – by Estimate Size

Estimation Domain	Estimate Size	Number of Estimates	Efficient Alloc, no Mins	Simple Alloc	Optimal Alloc 3	Optimal Alloc 4	Optimal Alloc 5	Optimal Alloc 6
National	Small	205	29.8%	13.2%	9.8%	13.2%	20.0%	15.1%
	Medium	408	37.7%	11.3%	8.1%	9.8%	17.9%	15.2%
	Large	204	44.1%	14.2%	6.9%	11.3%	13.7%	9.8%
State	Small	3,569	22.0%	17.8%	13.4%	12.2%	21.3%	28.7%
	Medium	7,199	27.1%	13.9%	9.0%	8.6%	18.1%	23.4%
	Large	3,605	42.9%	13.6%	7.9%	9.2%	14.4%	12.0%
NAICS4 Industry	Small	9,869	34.4%	28.1%	23.3%	23.4%	31.6%	33.8%
	Medium	49,337	25.2%	17.2%	11.7%	11.7%	21.6%	26.4%
	Large	25,214	31.4%	13.5%	7.9%	8.4%	17.9%	21.2%
MSA-BOS Area	Small	20,531	16.0%	18.6%	14.2%	13.0%	20.1%	27.0%
	Medium	82,415	16.7%	19.1%	12.5%	10.8%	17.5%	25.3%
	Large	43,303	27.8%	20.1%	10.2%	9.2%	15.8%	17.2%

NOTE: Summing the percentages across the columns do not result in 100 percent, because there can be more than one allocation that has the smallest RRMSE measure for a given estimate.

Table 5.2 shows that the efficient allocation with no minimums performs the best for national estimates for all three different sized estimates. For state and area estimates it appears that adding minimums helps the small and medium estimates, while the large estimates are still best under the efficient allocation. For the industry estimates, the efficient allocation and optimal allocation 6 perform about the same for the small and medium estimates. The large industry estimates are best under the efficient allocation.

6. Conclusion

There were several key findings from this research project. First, the performance of the optimization approach is dependent on the objective function used. We found unexpected negative consequences when using the seemingly reasonable objective functions in optimal allocations 1 and 2. These objective functions caused many of the detailed strata sample sizes to be adjusted down to almost zero, resulting in holes in the samples and the potential for extremely large sampling weights. If using an optimization approach for reconciling allocations, it is important to understand the effects the objective function will ultimately have on the final sample allocation.

Next, including minimum allocations appears to have very little effect on the performance of the estimates. This was a surprising result since we create the minimums to help the performance of the OES estimator. However, there is a silver lining in that these results show that we could impose minimums to the sub-nation estimation domains without a significant loss in the precision of the OES estimates. If the goal of adding minimums is to help with confidentiality rules or to ensure that every estimate has at least some survey data contributing to it (for modelled estimates), then this would be a positive result. In

future research we would like to test the robustness of this finding by testing different minimum allocations.

Lastly, we found that there were not significant improvements to the precision of the estimates when using the optimization approach versus the simple approach. This was also surprising since the simple approach appears to make less nuanced adjustments than the optimization approach. It seems that in our particular application the ability to adjust the allocation of each cell separately is not very important. There is some evidence in table 5.2 that the optimal allocations 5 and 6 outperform the simple allocation across each estimation domain, but only by a small amount. When looking at the overall distributions of the RRMSE measures in graphs 5.1 and 5.2, there appears to be no clear advantage of the optimization approach over the simple approach. Considering the potential of choosing an objective function which could have negative consequences and the added complexities of the optimization approach, the simple approach appears to be the better way of reconciling the minimum allocations with the efficient allocation for our particular application.

References

Hajek, J. 1958. Some contributions to the theory of probability sampling. Bull. Int. Stat. Inst. 36(3):127-134.

U.S. Bureau of Labor Statistics (2017). Handbook of Methods, Washington, DC. https://www.bls.gov/opub/hom/pdf/homch3.pdf .