Effect of Benchmark Cells Collapse Patterns on the National Compensation Survey Earnings Estimates

Dee A. Zamora[1], Chester H. Ponikowski[1], Jonathan J. Lisic[2]
[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 3160, Washington, D.C. 20212
[2]National Agricultural Statistics Service, 1400 Independence Ave., SW, Washington, D.C. 20250

Abstract
Benchmarking is used often in establishment surveys to adjust sample weights to match the current distribution of the population of interest. In the National Compensation Survey, the weight of each establishment in the sample is adjusted to match the distribution of current employment by industry from the Quarterly Census of Employment and Wages program. The process involves calculating a benchmark factor for each cell and multiplying the establishment weight by the calculated factor. In cases where there are fewer than three responding sample establishments or the factor is larger than 4.00, two or more cells are collapsed. The question is which cells should be collapsed so that the effect on the mean square error is minimized. This paper presents the current collapse pattern and several other collapse patterns and evaluates their impact on mean square error of earnings estimates.

Key Words: bias, variance, weighting cells

1. Introduction

In many large establishment surveys, a considerable amount of time can pass from the time when a sample is selected to when estimates are generated. During this period there is a potential for shifts in employment distribution to occur due to uneven growth or decline in industry employment. Benchmarking or post-stratification helps to correct for these shifts in industry employment and also can offset sampling variability. In this process, benchmark weights are created such that the sum of final weights will correspond to a targeted control number usually obtained from an independent, outside source.

In the National Compensation Survey (NCS), the estimation process occurs well after the sample selection process, when economic conditions may have changed. The NCS employs the benchmarking process to create factors that adjust the sampling weight of each sample establishment so that employment estimates are reflective of the economy of the reference period.

During the benchmarking process for the locality wage outputs of the NCS, benchmark factors are calculated for each industry cell. When the benchmark factor for an industry cell is larger than 4.00 or when there are fewer than three contributing establishments, the cell is combined or collapsed with one or more other cells to form a new data cell. Cell

collapsing is used as a way to minimize variance but with the attempt not to have an adverse effect of increased bias. A maximum factor is capped at 4.00 in order to prevent any one response from being overrepresented in the estimate. Benchmark factors are one of several adjustment factors that are applied before estimation begins. The NCS also applies factors that address unit non-response, item non-response, and aberrant collection situations. A more detailed explanation of these factors is provided in Chapter 8 of BLS Handbook of Methods.

A number of articles have been written discussing the uses and benefits of post-stratification (Cervantes and Brick 2009; Jayasuriya and Valliant 1995; Holt and Smith 1979; Zhang 2000). More recently, Kim, Li and Valliant offered findings from their research specific to cell collapsing in post-stratification. In their research, they looked at the effects on bias and variance from cell collapsing (Kim, Li, and Valliant 2007).

This paper explores whether the choice in cell collapsing or collapse patterns impacts the accuracy of estimates. The current cell collapse patterns used during the benchmarking process are compared to alternative patterns. The basis of our evaluation is the comparison of the root mean square errors (RMSE) of the estimates produced from each of the tested collapse patterns. Section 2 provides a brief description of the NCS. Section 3 presents empirical analysis and results. Conclusion and options for further research are presented in section 4.

2. Description of the National Compensation Survey

The National Compensation Survey (NCS) is an establishment survey of wages and benefits conducted by the Bureau of Labor Statistics (BLS). From the survey, three general survey outputs are produced: employment cost data, employee benefits data, and locality and national wage data. The locality and national wage data are published annually and provide occupational wages for a sample of localities, census divisions, and the nation as a whole. The survey covers state and local governments and private sector industries, except for agriculture and private households.

The locality wage outputs from the NCS provide wage estimates for more than 800 occupations of roughly 80 metropolitan and non-metropolitan localities. The estimates include wage data by industry, occupational group, full-time and part-time status, union and nonunion status, establishment size, and job level. Wage data are presented for total workers as well as separated by private industry and state and local government. Such information allows users to compare wages of different occupations across the nation.

The NCS uses the BLS Quarterly Census of Employment and Wages (QCEW) as its sampling frame from which we used the administrative data of establishments for this study. Data for the QCEW are obtained from the State Unemployment Insurance files of establishments that are collected by the individual state agencies. The files include the descriptive variables, monthly employee count, and quarterly total wages of all employees for each establishment.

The integrated NCS sample is made up of five rotating replacement sample panels. Each of the five samples panels are in the sample for five years after which they are replaced by a panel that is selected from a more current frame. The NCS employs a three-stage stratified design with probability proportionate to employment sampling at each stage. At

the first stage a probability sample of areas is selected; the second stage is a probability sample of establishments within the sampled areas; and in the third stage a probability sample of occupations is drawn within the sampled areas and establishments.

The most recent NCS area sample consists of 152 areas which were selected using the 2003 OMB set of area definitions. The area definitions define a set of Core Based Statistical Areas (CBSA) and designate the remaining geographical areas as outside CBSA counties. The outside CBSA areas for NCS sampling purposes are usually clusters of adjacent counties, not single counties. A more detailed description of the NCS sample design is given in Chapter 8 of Handbook of Methods and in the 2005 ASA paper, "Phase-in of the Redesigned National Compensation Survey Area Sample."

Each establishment in the QCEW is assigned an industry classification based on definitions provided by the North American Industry Classification System (NAICS) and designated as having a private, local or state government ownership. The classification groups of the NAICS are based on industries with similar production orientations (process, technology, functions) as opposed to a market based system. It is thought that this allows for a more comparable basis of economic and statistical data.

The current collapse pattern of the locality wage program approximately follows the ordering of industries under NAICS. Industry cells are grouped with industry cells that are in close proximity to their NAICS code. For example, "arts, entertainment, and recreation" (NAICS 71), "accommodation and food services" (NAICS 72), and "other services" (NAICS 81) form one collapse pattern. Similarly, the "information" (NAICS 51), "finance and insurance" (NAICS 52), and "real estate and rental and leasing" (NAICS 53) form another collapse pattern. The notable deviations from this sequential methodology are the grouping of "mining" and "construction" industry sectors (NAICS sectors 21 and 23) and the combination of "utilities" and "transportation and warehousing" (NAICS sectors 22 and 48).

The current collapse pattern is presented in Table 1, below. For each of the industry cells, which are defined in the "NAICS Code" and "Industry Name" columns, the industry cell(s) with which it collapses is given in the "Current Pattern" column. For example, if industry cell 21A has less than three reporting establishments or has a benchmark weight greater than 4.00, then it is collapsed with industry cell 23A to form a new industry cell and a new benchmark weight is calculated for this new cell.

After collapsing is done, it is possible that the conditions of three responding establishments and a factor no greater than 4.00 may not be met. In this case, a second level of collapsing would occur; the collapse pattern is four large groups of industry cells. In this study, we limited the collapsing to the first level.

Table 1. Collapse Patterns by Industry Cell

Industry Cell	NAICS Code	Industry Name	Current Pattern
21A	21	Mining	21A, 23A
22A	22	Utilities	22A, 48A
23A	23	Construction	21A, 23A
31A	31-33 (excl. 336411)	Manufacturing	31A, 31B
31B	336411	Aircraft Manufacturing	31A, 31B
42A	42	Wholesale Trade	42A, 44A
44A	44-45	Retail Trade	42A, 44A
48A	48-49	Transportation and Warehousing	22A, 48A
51A	51	Information	51A, 52A, 52B, 53A
52A	52 (excl. 524)	Finance (Rest of)	51A, 52A, 52B, 53A
52B	524	Insurance	51A, 52A, 52B, 53A
53A	53	Real Estate, Renting, Leasing	51A, 52A, 52B, 53A
54A	54	Professsional, Scientific Technical	54A, 55A, 56A
55A	55	Management of Companies and Enterprises	54A, 55A, 56A
56A	56	Admin., Support, Waste Management	54A, 55A, 56A
61A	61 (excl. 6111-6113)	Educational Services (Rest of)	61A, 61B, 61C
61B	6111	Elementary and Secondary Schools	61A, 61B, 61C
61C	6112, 6113	Junior Colleges, Colleges and Universities	61A, 61B, 61C
62A	62 (excl. 622, 623)	Healthcare, Social Assistance (Rest of)	62A, 62B, 62C
62B	622	Hospitals	62A, 62B, 62C
62C	623	Nursing and Residential Care Facilities	62A, 62B, 62C
71A	71	Arts, Entertainment, Recreation	71A, 72A, 81A
72A	72	Accomodation and Food Services	71A, 72A, 81A
81A	81 (excl. 814)	Other Services (excl. Public Administration)	71A, 72A, 81A

3. Empirical Analysis and Results

First, we selected test areas based on their incidence of industry cells that required collapsing and selected six locality areas to test: Buffalo, Huntsville, Louisville, Palm Bay, San Antonio, and Tallahassee. As displayed below in Table 2, we chose three alternative collapse patterns to test against the current collapse pattern. Alternative 1 is based on pairs of cells with similar average wage under the reasoning that the industry cells that are similar in average wage would yield the estimates with the smallest error. Alternatives 2 and 3 are two sets of collapse patterns based on their sequential industry cell order. Alternative 2 was created by forming pairs of cells starting with the first ordered cell (i.e., 21A and 22A collapse together); in Alternative 3 we formed pairs starting with the second ordered cell (i.e., 21A does not collapse with any other cell, 22A and 23A collapse together).

Table 2. Alternative Collapse Patterns by Industry Cell

	Alternative Collapse Patterns Tested		
Industry Cell	**Alternative 1**	**Alternative 2**	**Alternative 3**
21A	21A, 55A	21A, 22A	21A
22A	22A, 52B	21A, 22A	22A, 23A
23A	23A, 62A	23A, 31A	22A, 23A
31A	31A, 31B	23A, 31A	31A, 31B
31B	31A, 31B	31B, 42A	31A, 31B
42A	42A, 51A	31B, 42A	42A, 44A
44A	44A, 72A	44A, 48A	42A, 44A
48A	48A, 62B	44A, 48A	48A, 51A
51A	42A, 51A	51A, 52A	48A, 51A
52A	52A, 54A	51A, 52A	52A, 52B
52B	22A, 52B	52B, 53A	52A, 52B
53A	53A, 56A	52B, 53A	53A, 54A
54A	52A, 54A	54A, 55A	53A, 54A
55A	21A, 55A	54A, 55A	55A, 56A
56A	53A, 56A	56A, 61A	55A, 56A
61A	61A, 71A	56A, 61A	61A, 61B
61B	61B, 61C	61B, 61C	61A, 61B
61C	61B, 61C	61B, 61C	61C, 62A
62A	23A, 62A	62A, 62B	61C, 62A
62B	48A, 62B	62A, 62B	62B, 62C
62C	62C, 81A	62C, 71A	62B, 62C
71A	61A, 71A	62C, 71A	71A, 72A
72A	44A, 72A	72A, 81A	71A, 72A
81A	62C, 81A	72A, 81A	72A, 81A

For this study, we used the NCS establishment sample that had data collected in 2005. We obtained the establishment-level employment and wage information from the 2005 QCEW data for the "usable" establishments (or viable, in-scope, and responding establishments at initiation); sampled establishments that could not be identified in the administrative data were excluded from the study.

We decided on December, 2008 as the reference period to generate estimates, which we thought would provide enough of a time difference to observe an employment shift. For each industry cell, we calculated the benchmark factor. The numerator of the benchmark factor was the target control number that was calculated using the employment information from the December, 2008 administrative data; the denominator was the weighted survey employment that was calculated using the employment information from the 2005 administrative data and the establishment weights of the usable establishments from our sample.

Based on the cell collapsing criteria, we identified the cells requiring collapsing which are presented in Table 3, presented below. Of the 144 area-industry cells, 49 cells required collapsing, all of which did not meet the requirement of three contributing units. Six of the 49 cells had benchmark factors greater than 4.00. For the 24 industry cells, 10 cells required collapsing in at least three test areas, five required collapsing in one or two areas and nine cells did not require any collapsing. Of the tested areas, Palm Bay and Tallahassee had more collapsing performed than the other areas and this is explained by the smaller sample sizes for these areas.

Table 3. Industry Cells Requiring Collapsing by Area

Industry Grouping	Industry Cell	Buffalo	Huntsville	Louisville	Palm Bay	San Antonio	Tallahassee
Construction and Mining	21A	x	x	x	x	x	x
	23A						
Manufacturing	31A						
	31B	x	x		x	x	
Trade, Transportation, and Utilities	22A	x	x		x	x	x
	42A				x		x
	44A						
	48A				x		
Information	51A				x		x
Financial Activities	52A						
	52B				x		
	53A	x			x		x
Professional and Business Services	54A						
	55A		x	x	x	x	x
	56A						x
Education and Health Services	61A	x			x	x	x
	61B			x	x		x
	61C	x	x	x	x		x
	62A						
	62B						
	62C		x		x		x
Leisure and Hospitality	71A	x		x		x	x
	72A						
Other Services	81A						
All Private Industry	Total	7	6	5	13	6	12

For industry cells requiring collapsing, we calculated the benchmark factor for the collapsed cell. The numerator was the target control number, which was the sum of the December, 2008 employments of the industry cells forming the collapsed cell. The denominator was the sum of the survey weighted employments of the industry cells forming the collapsed cells. We calculated the benchmark factors using the current collapse pattern and the three test patterns. We observed that many of the benchmark factors were between 0.8 and 2.0 and that the factors did not differ much among the patterns for an industry cell. The largest differences were for the six industry cells that had a benchmark factor greater than 4.00 before collapsing.

After calculating the benchmark factors, we used the factors to adjust the sample establishment weight in computing the average monthly earnings for major private industry groups and for all private industry. The major industry groupings are defined by their industry cells in Table 3. The monthly average was computed as

$$\frac{\Sigma_i \Sigma_j \mathit{Monthly\ Average\ Wage}_{ij} * \mathit{Employment}_{ij} * \mathit{Establishment\ Weight}_{ij} * BF_i}{\Sigma_i \Sigma_j \mathit{Employment}_{ij} * \mathit{Establishment\ Weight}_{ij} * BF_i}$$

Where *BF* is the benchmark factor, *i* is the industry cell, and *j* is the establishment. For a given industry grouping, we calculated five average monthly earnings based on the benchmark factors we calculated under the scenarios of no collapsing, the current collapse pattern, and the three alternative collapse patterns.

Since the same sampled units were used in the earnings estimates for a survey area, the monthly average wage, employment and establishment weight for each sampled unit remained the same in the computation, and the change in the estimates was isolated to the benchmark factors from a given collapse pattern.

Using the administrative data of the reference period, we calculated the average monthly earnings from the frame to arrive at population values for the major private industry groups and all private industry. We used these population values to calculate the bias of the area-industry group estimates. The variances were computed using the balanced repeated replication (BRR) methodology. A description of this methodology can be found in Wolter (2007).

After calculating the bias and variance, we then calculated the root mean square error (RMSE) for all the estimates. Table 4 shows the sample estimates of the average monthly earnings of all private industry and their RMSE for each collapse pattern. The collapse pattern with the lowest RMSE for an area estimate is shown in italics. For each area estimate, the RMSE is lower under one of the collapse patterns than if no collapsing had been done, although the collapse pattern that resulted in the lowest RMSE was not consistent for the areas.

Table 4. Average Monthly Earnings Estimates and RMSE for All Private Industry by Survey Area and Collapse Pattern

Area	Collapse Pattern	Sample Estimate	RMSE
Buffalo	No Collapse	2,892.77	422.29
	Current	2,867.81	448.38
	Alternative 1	2,883.71	431.04
	Alternative 2	2,903.98	411.02
	Alternative 3	2,916.08	*399.39*
Huntsville	No Collapse	3,214.76	618.91
	Current	3,269.42	*568.24*
	Alternative 1	3,218.66	615.12
	Alternative 2	3,224.84	609.02
	Alternative 3	3,213.54	619.98
Louisville	No Collapse	2,928.00	544.57
	Current	2,946.85	527.39
	Alternative 1	2,928.86	543.75
	Alternative 2	2,917.07	556.30
	Alternative 3	2,954.68	*519.04*
Palm Bay	No Collapse	3,071.47	608.75
	Current	3,107.70	*572.63*
	Alternative 1	3,062.34	618.41
	Alternative 2	3,061.14	619.12
	Alternative 3	3,095.25	584.34
San Antonio	No Collapse	2,916.03	450.18
	Current	2,885.65	479.83
	Alternative 1	2,919.16	*449.36*
	Alternative 2	2,869.69	495.63
	Alternative 3	2,885.53	480.56
Tallahassee	No Collapse	2,678.82	526.37
	Current	2,833.15	*392.69*
	Alternative 1	2,728.75	480.62
	Alternative 2	2,708.97	501.95
	Alternative 3	2,717.71	491.31

Looking at the estimates and RMSE for the other area-industry groupings, we observed a lack of a consistent pattern for the lowest RMSE either by industry grouping or area. That is, no area or industry grouping seemed to produce estimates with the lowest RMSE from the same collapse pattern. In Table 5, we indicate which collapse pattern resulted in the lowest RMSE by area-industry estimate. There are some area-industry estimates where more than one pattern had the lowest RMSE. In some cases, this occurred because no collapsing was done and the estimates for all patterns were the same. There were also cases where there was only one industry cell with sampled units for an industry grouping; in these cases, the estimates for the industry grouping are the same. The results do appear to support collapsing over not collapsing. In most of the area-industry estimates, the estimates that underwent collapsing had the lowest RMSE.

Table 5. Collapse Pattern with Lowest RMSE by Area-Industry Estimate

Area	Estimate	No Collapse	Current Collapse	Alternative 1	Alternative 2	Alternative 3
Buffalo	All Private					x
	Construction & Mining		x			
	Education & Health Services		x			
	Financial Activities		x			
	Leisure & Hospitality			x		
	Professional & Business Services					x
	Trade, Transportation, and Utilities				x	
Huntsville	All Private		x			
	Construction & Mining	x				
	Education & Health Services					x
	Financial Activities			x		
	Leisure & Hospitality				x	
	Professional & Business Services		x			
	Trade, Transportation, and Utilities				x	
Louisville	All Private					x
	Construction & Mining	x				x
	Education & Health Services		x			
	Financial Activities	x	x	x	x	x
	Leisure & Hospitality	x				
	Professional & Business Services					x
	Trade, Transportation, and Utilities	x	x	x		x
Palm Bay	All Private		x			
	Construction & Mining	x	x	x	x	x
	Education & Health Services			x		
	Financial Activities				x	
	Leisure & Hospitality				x	
	Professional & Business Services			x		
	Trade, Transportation, and Utilities			x		
San Antonio	All Private			x		
	Construction & Mining	x				
	Education & Health Services		x			
	Financial Activities			x		
	Leisure & Hospitality				x	
	Professional & Business Services			x		
	Trade, Transportation, and Utilities	x				
Tallahassee	All Private		x			
	Construction & Mining	x	x	x	x	x
	Education & Health Services					x
	Financial Activities	x				
	Leisure & Hospitality	x				
	Professional & Business Services		x			
	Trade, Transportation, and Utilities			x		

4. Conclusion and Issues for Further Research

In this study, we have attempted to assess what impact cell collapsing during benchmarking may have on estimates. We began by looking at areas that historically have had collapsing performed and used administrative data to calculate monthly average wages under various collapse patterns for actual private samples selected for these areas. We used the root mean square error to evaluate collapse patterns. From the administrative data, we were able to calculate population values which we then could use to calculate the bias of our estimates. We computed the variance using BRR.

From our evaluation, our results suggest that collapsing improves estimates over not collapsing. We found no evidence to favor one collapse pattern over another tested pattern. Moreover, no tested pattern performed better for a particular area or industry grouping estimate.

This study was limited to a single sample of six areas for one time period. If we continue this research, there are several ways to extend the research. Using the same method, we

could study the collapse patterns for these areas or other areas with different collection and reference periods.

We would like to explore the decision to use industry as a basis for forming benchmark cells and collapsing. In our study, the benchmark cells and collapsing were based on industry; however, we would like to study benchmarking and collapsing on a different basis such as employment size classes.

Another option would be to use simulated samples. Rather than using a single actual sample that was selected, we could pursue selecting multiple samples from a single frame and calculate estimates and benchmark factors for each sample based on expected response rates. A benefit to conducting simulated samples is that it would alleviate concerns that estimates produced from one sample are more subject to errors in sampling variability.

By pursuing any of these research options, we hope to either provide more support for our initial results or more definitive conclusions on the impact of collapse pattern selections.

References

BLS Handbook of Methods (2009). Washington, D.C.: Bureau of Labor Statistics, Chapter 8. http://www.bls.gov/opub/hom/homch8_a.htm.

Bureau of Labor Statistics: National Compensation Survey, Summary. http://www.bls.gov/ncs/summary.htm

Cervantes, I.F. and Brick, J.F. (2009). Efficacy of Poststratification in Complex Sample Design. *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association. 4642-4655.

Holt D. and Smith, T.M.F. (1979). Post Stratification, *Journal of the Royal Statistical Society*, 142, 33-46.

Jayasuriya, B. and Valliant, R. (1995). An Application of Regression and Calibration Estimation to Post-Stratification in a Household Survey. *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association.

Kim, J.J., Li, J., and Valliant, R. (2007). Cell Collapsing in Poststratification. *Survey Methodology, 33, 139-150.*

Office of Management and Budget (2007). *North American Industry Classification System: United States, 2007.* Washington, D.C.: OMB.

Tehonica, J., Ernst, L.R., and Ponikowski, C.H. (2005). Phase-in of the Redesigned National Compensation Survey Area Sample. *Proceedings of the Survey Research Methods Section*, Alexandria, VA.: American Statistical Association.

Wolter, Kirk M. (2007). *Introduction to Variance Estimation.* New York: Springer Science+Business Media, LLC.

Zhang, Li-Chun (2000). Post-Stratification and Calibration- A Synthesis. *The American Statistician*, 54, 178-184.

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