SMALL AREA RESEARCH FOR THE OCCUPATIONAL EMPLOYMENT STATISTICS SURVEY

Bogong T. Li, Stephen M. Miller,

U.S. Bureau of Labor Statistics

Key Words: Synthetic estimators, Finite population.

Abstract:

The Occupational Employment Statistics (OES) Survey is a yearly mail survey designed to produce estimates of employment and wages for more than 700 occupations in the U.S. The OES sample is stratified by geographic area, economic activity, and employment size class; with geographic area defined by State, Metropolitan Statistical Area (MSA) and balance of State area, and economic activity defined by 3-digit Standard Industrial Classification (SIC) codes. While the sample is designed to produce reliable design-based estimates for large geographic areas, our research investigates synthetic estimators for smaller geographic areas, such as at the county level.

1. The Occupational Employment Statistics Survey

The Occupational Employment Statistics (OES) survey program at the U.S. Bureau of Labor Statistics (BLS) is a yearly mail survey designed to produce estimates of employment and wages for specific occupations by geographic areas and by industry. Estimates based on geographic areas are available at the National, State, and Metropolitan Statistical Area (MSA) levels. Estimates of occupational employment (OE) and wages for over 400 industries are available at the national level. The OES samples approximately 400,000 establishments per year, collecting data on wage and salary of workers in nonfarm establishments for over 700 occupations.

The OES program, as a Federal-State cooperative program, provides information heavily utilized by both federal and local policy makers. Increasingly States are interested in producing estimates for additional geographic areas. Many of the additional areas of interest are at the county level.

Estimation of OE and wages are available at a basic level defined by State, MSA, and 3-digit SIC and at aggregates of the basic level. The weighted sample employment, at the basic estimation level, is adjusted to the total employment derived from BLS's Longitudinal Database. Any estimates below the basic level, such as at the sub-MSA level, are not controlled, and therefore are subject to bias and a high level of sampling variability. We consider in this study three synthetic estimators to produce sub-MSA OE estimates which may provide higher statistical accuracy.

2. Synthetic Estimators for Local Occupational Employment

2.1 Notation

This section lists the notation that will be used throughout the paper.

- E total employment, e.g. $\mathbf{E}_{d,g}$ is the total employment of population group g in small domain d; $\mathbf{E}_{d,.}$ is the total employment in small domain d.
- $\mathbf{P}_{,,o}$ percentage of OE for occupation o, e.g., $\mathbf{P}_{d,g,o}$, is the percentage

The views expressed in this article are those of the authors and do not constitute the policy of the Bureau of Labor Statistics. The authors can be contacted at: U.S. Bureau of Labor Statistics, Postal Square Building, Suite 1950, 2 Massachusetts Ave. NE. Washington, DC 20212-000, or E-Mail: li_t@bls.gov; miller_s@bls.gov.

of OE for occupation o among establishment group g and small domain d. $\mathbf{P}_{,,o}$ are not known at any level. We need to estimate them through survey data, usually by a ratio estimator.

$\mathsf{X}_{,o}$ total OE at various levels, e.g. $\mathsf{X}_{i,o}$, is the total OE of occupation o in establishment i; $\mathbf{X}_{d,o}$, is the total OE in small domain d. Only some $\mathsf{X}_{i,o}$ are available through survey data. Our goal in this study is to estimate the total OE, $\mathbf{X}_{d,o}$ at the sub-MSA, or small domain level.

w_i the sampling weight for establishment i, which in practice should include nonresponse adjustments.

d index for an estimation domain, or small domain, at the sub-MSA level, $d = 1, \ldots, D$.

g index for an establishment group. Establishments grouped by g are not grouped by geography in general. In particular, here index g groups all establishments in a State belonging to the same SIC2-digit industry, $g = 1, \ldots, G$.

h index of an establishment group defined by MSA/SIC3-Industry/Size-Class. This is a finer division of the establishments than those defined by g, $h = 1, \ldots, H$.

i establishment index.

o occupation index.

$\{\}_d$ the group of establishments belonging to small domain d.

$\{\}_g$ the group of establishments belonging to an SIC2 industry group g.

$\{\}_h$ the group of establishments belonging to an MSA/SIC3-Industry/Size-Class cell h.

Horvitz-Thompson type estimators are used to estimate total employment, OE or percentages of OE at the appropriate level of aggregation.

2.2 Synthetic Estimation Method

The OES local OE synthetic estimator assumes a fixed sample design over a finite population of establishments. Currently the OES program provides OE estimates at the MSA/SIC3/Size-Class level. In order to obtain the estimate at a sub-MSA level, that is, at a small domain indexed by d, the OES further assumes any OE proportion, which is the particular share of occupation o among total employment within a group h is consistent across all small domains. Therefore we estimate $\mathbf{X}_{d,o}$, the total OE in small domain d by

$$\widehat{\mathsf{X}}^{(1)}_{d,o} = \sum_{h=1}^{H} \mathbf{E}_{d,h} \cdot \widehat{P}_{\cdot,h,o} \qquad (1)$$

where $\widehat{P}_{\cdot,h,o} = \frac{\sum_{i\in\{\}_h} w_i \cdot \mathsf{X}_{i,o}}{\widehat{\mathrm{E}}_{\cdot,g}}$ and $\widehat{\mathrm{E}}_{\cdot,g} = \sum_{o=1}^{O} \sum_{i\in\{\}_h} w_i \cdot \mathsf{X}_{i,o}$. This is our first synthetic estimator to consider, we call it Estimator 1.

Estimator 1 provides reasonable estimates given the homogeneity assumption is satisfied. We will show in the Appendix this estimator as well as Estimator 2 are biased if this assumption is violated. Since this group is relatively finely defined, the chance of an establishment from a county where its industry group is rare has smaller chance of being included in the sample, therefore it is possible we can not estimate for some counties. Also for synthetics estimators in general, extremely fine grouping reduces efficiency. Empirical analysis in the past does not show improvement over finely grouped populations similar to OES. One of the early references is that of National Center for Health Statistics ([4] 1968) and Gonzales ([2] 1973).

Estimator 2 groups establishments within a State by SIC2-digit industry division, assuming establishments belonging to the same SIC2-digit industry division share the same OE distribution. In our notation, the share of a particular OE among total employment in the SIC2-digit industry division is consistent across all ge-

ographic locations, or small domains,

$$\mathbf{P}_{\cdot,g,o} = \mathbf{P}_{d,g,o} \quad \text{for all } g,\ g = 1, \ldots, G\,.$$

In this case, similar to Estimator 1, we can obtain an estimate for $\mathbf{X}_{d,o}$ through a ratio estimator,

$$\widehat{\mathsf{X}}^{(2)}_{d,o} = \sum_{g=1}^{G} \mathbf{E}_{d,g} \cdot \widehat{P}_{\cdot,g,o} \tag{2}$$

where $\widehat{P}_{d,g,o} = \frac{\sum_{i\in\{\}_g} w_i \cdot \mathsf{X}_{i,o}}{\widehat{\mathrm{E}}_{\cdot,g}}$ and $\widehat{\mathrm{E}}_{\cdot,g} = \sum_{o=1}^{O}\sum_{i\in\{\}_g} w_i \cdot \mathsf{X}_{i,o}$. However estimator $\widehat{\mathsf{X}}^{(2)}_{d,o}$ is biased if the homogeneity assumption is violated. This is because the OE total is $\mathbf{X}_{d,o} = \sum_{g=1}^{G} \mathbf{E}_{d,g} \cdot \mathbf{P}_{d,g,o}$, and the bias of $\widehat{\mathsf{X}}^{(2)}_{d,o}$ is

$$\begin{aligned}\text{bias} &= \mathbf{X}_{d,o} - \mathbb{E}(\widehat{\mathsf{X}}^{(2)}_{d,o}) \\ &= \sum_{g=1}^{G} \mathbf{E}_{d,g}(\mathbf{P}_{d,g,o} - \mathbf{P}_{\cdot,g,o})\,. \end{aligned} \tag{3}$$

This bias is zero if $\mathbf{P}_{d,g,o} = \mathbf{P}_{\cdot,g,o}$ for every establishment group g, $g = 1, \ldots, G$. This is hardly true in reality.

However we can construct a synthetic estimator with a bias correction, $\widehat{\mathsf{X}}^{(2)}_{d,o} + \text{bias}$, if we can estimate the bias in some way. We substitute ratio estimators $\widehat{P}_{\cdot,g,o}$,

$$\widehat{P}_{d,g,o} = \frac{\sum_{i\in\{\}_d\cap\{\}_g} w_i \cdot \mathsf{X}_{i,o}}{\mathbf{E}_{d,g}} \quad \text{and}$$

$$\widehat{\mathrm{E}}_{d,g} = \mathbf{E}_{d,\cdot}\,\frac{\widehat{\mathrm{E}}_{d,g}}{\widehat{\mathrm{E}}_{d,\cdot}}$$

for $\mathbf{P}_{\cdot,g,o}$, $\mathbf{P}_{d,g,o}$ and $\mathbf{E}_{d,g}$ in the bias expression (3) to estimate the bias. This leads to the third estimator which we call Estimator 3,

$$\begin{aligned}\widehat{\mathsf{X}}^{(3)}_{d,o} &= \widehat{\mathsf{X}}^{(2)}_{d,o} + \widehat{\text{bias}} \\ &= \widehat{\mathsf{X}}^{(2)}_{d,o} + \sum_{g=1}^{G} \widehat{\mathrm{E}}_{d,g}(\widehat{P}_{d,g,o} - \widehat{P}_{\cdot,g,o})\,. \end{aligned}$$

After some rearrangement of terms in the above expression (see in Appendix) we have

$$\widehat{\mathsf{X}}^{(3)}_{d,o} = \mathbf{E}_{d,\cdot}\left[\widehat{P}_{d,\cdot,o} + \sum_{g=1}^{G}\left(\frac{\mathbf{E}_{d,g}}{\mathbf{E}_{d,\cdot}} - \frac{\widehat{\mathrm{E}}_{d,g}}{\widehat{\mathrm{E}}_{d,\cdot}}\right)\widehat{P}_{\cdot,g,o}\right] \tag{4}$$

In this form the estimation of a particular OE in small domain d is expressed as a proportion of total employment of small domain d, with the expression in the brackets as an estimate of the percentage of OE in small domain d.

3. Simulation Study and Results

In this section we evaluate the performance of these three estimators with a simulation study. This is necessary because of the unavailability of complete information on every establishment in US. The simulation draws samples from an artificial establishment population equipped with occupational employment characteristics very similar to the actual population. The three proposed estimators are computed based on the random samples and then compared to each other and to the simulated population values.

The simulated population is produced based on a large U.S. State sample. Establishment characteristics such as the distribution of total establishment employment, the proportion of OE to total employment, and the distribution of OE across sub-MSA area, etc. are closely modeled after the large State sample. This simulated population contains six counties in two MSAs with a total of one thousand establishments. The employment types include a total of fifty occupational categories. The establishment sizes are in seven categories with employees ranging from one to three thousand. A comparison of the unit employment distribution between the simulated population and the large U.S. State sample reveals a close similarity between the simulated and true sample population in terms of the characteristics of the labor force. The sampling design stratifies the one thousand establishments by three-digit industry code and size class code. Establishment employing 250 employees or more are sampled with certainty. Establishments employing fewer than 250 employees and more than two employees are sampled with probability proportional to the size class employment within each three-digit industry. Establishments with only one employee are included in the sample only if it is the only establishment in the SIC3/Size-Class

	Estimator 1	Estimator 2	Estimator 3
Mean Standard Error	3.049	12.172	2.508
Mean Coefficient of Variation	0.090	0.096	0.087
Average Bias	-2.514	-2.534	*-0.104*
Average Relative Bias	-0.027	-0.087	0.051
Overall Error Rate	0.059	0.137	0.05

Table 1: OE Estimates for Six Counties

	Estimator 1	Estimator 2	Estimator 3
95% Confidence Interval Nominal Coverage	0.942	0.955	0.968
Standardized Error Confidence Bound Estimate	(-.513, .39)	(-.32, .14)	(-.45, .47)
Bias Confidence Bound Estimate	(-39.2, 29.9)	(-48.6, 20.8)	(-34.9, 36.1)

Table 2: 20^{th} Occupational Employment (OE) Estimate in County One

cell. Within each SIC3/Size-Class cell, establishments are systematically selected into the sample through a single random start.

Each time we draw a sample, we calculate the employment for every occupation within each county separately using three estimators. There are a total of 263 estimates for 6 counties from each sample. Each sample contains 250 establishments. Table 1 summaries the overall estimation results. All measures are averages from the 200 independent random samples. Table 2 lists the estimation summary for a particular occupation randomly selected from a local area, in this case, the No. 20 occupation in county one.

The mean standard error for Estimator 2 is much higher than that of Estimator 1 and 3, the larger grouping of similar establishments increased the variability. Estimator 3 has about half of the bias of the other two estimators. The reduction is possibly the result of the additional bias adjustment term. The mean CVs of all three estimators are all similar. The Overall Error Rate (OER) which is define as

$$\text{OER} = \frac{||E(\widehat{\mathbf{X}}) - \mathbf{X}||}{||\mathbf{X}||}$$

is significantly higher for Estimator 2. OER measures the percentage error of the estimated vectors when compared to the true value. In addition, all three estimators have strong confidence interval coverage.

Another interesting observation is that although we would expect as the sampling fraction (SF) (percentage sample over the size of the sampling frame) increases the error rate of the estimators would decrease, the OER of Estimator 2 decreases much more slowly than the other two estimators: see Figure 1. Estimator 2 has smaller OER only when SF < 0.05 while Estimator 3 clearly does better when SF > 0.2 though its advantage relative to Estimator 1 is small. The phenomena could be the consequence of (1) the complicating effect of the sampling design and estimation method, (2) the effect of small cells on estimation in Estimator 1 and 3, (3) the computing program's handling of empty cells. Further investigation is clearly needed.

4. Conclusion

We conclude that the three synthetic estimators have similar confidence interval coverage and similar CV's. However, Estimator 3 does seem to have smaller estimation bias, while Estimator 2 has the largest mean standard error.

Future research will focus on improvement of the simulation population so it resembles the true population more closely. In addition, we will apply these methods to other real U.S. State

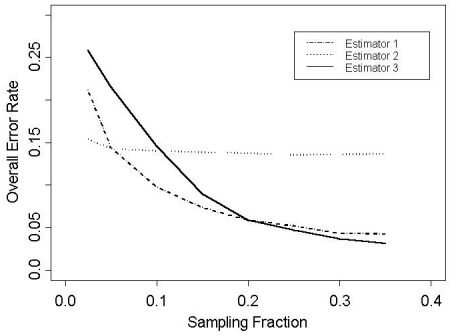

Figure 1: Overall Error Rates (OER) decrease as sampling fraction increases at different rates

establishment populations as well as incorporate additional Small Area Models.

Appendix

Depending on the level of grouping, the OE for small domain d under a finite population setting is expressed either as $\mathbf{X}_{d,o} = \sum_{h=1}^{H} \mathbf{E}_{d,h} \cdot \mathbf{P}_{\cdot,h,o}$ if the population grouping is defined at MSA/SIC3/Size-Class level or as $\mathbf{X}_{d,o} = \sum_{g=1}^{G} \mathbf{E}_{g,h} \cdot \mathbf{P}_{\cdot,g,o}$ if defined at MSA/SIC2/Size-Class level. Synthetic Estimator 1, $\widehat{\mathsf{X}}^{(1)}_{d,o}$ is approximately unbiased because

$$\mathbb{E}(\widehat{\mathsf{X}}^{(1)}_{d,o}) = \mathbb{E}\Big(\sum_{h=1}^{H} \mathbf{E}_{d,h} \cdot \widehat{P}_{\cdot,h,o}\Big) \simeq \sum_{h=1}^{H} \mathbf{E}_{d,h} \cdot \mathbf{P}_{\cdot,h,o} = \mathbf{X}_{d,o}\,,$$

the expectation is taken over all simple random samples selected without replacement over all establishments, given conditions to ensure the relative unbiasedness of ratio estimators, see Cochran ([1]).

Estimator $\widehat{\mathsf{X}}^{(2)}_{d,o}$ is biased unless the bias is zero:

$$\text{bias}(\widehat{\mathsf{X}}^{(2)}_{d,o}) = \sum_{g=1}^{G} \mathbf{E}_{d,g}(\mathbf{P}_{d,g,o} - \mathbf{P}_{\cdot,g,o}) = 0\,.$$

However this is true only if $\mathbf{E}_{d,g}(\mathbf{P}_{d,g,o} - \mathbf{P}_{\cdot,g,o}) = 0$ for all g, $g = 1, \ldots, G$ which is not true in reality. Therefore a bias may occur by using $\widehat{\mathsf{X}}^{(2)}_{d,o}$. We can adjust the estimate by adding an estimated bias, $\widehat{\text{bias}}(\widehat{\mathsf{X}}^{(2)}_{d,o})$ to $\widehat{\mathsf{X}}^{(2)}_{d,o}$. It leads to our Estimator 3, $\widehat{\mathsf{X}}^{(3)}_{d,o}$. This estimator improves upon $\widehat{\mathsf{X}}^{(2)}_{d,o}$ by adding a bias correction.

Next we show that equation (4) is true.

Under the finite population setting, we use the following estimators to estimate the percentage of OE at various population group levels:

$$\widehat{P}_{d,\cdot,o} = \frac{\sum_{i\in\{\}_d} w_i \cdot \mathsf{X}_{i,o}}{\widehat{\mathrm{E}}_{d,\cdot}}\,,$$

$$\widehat{P}_{\cdot,g,o} = \frac{\sum_{i\in\{\}_g} w_i \cdot \mathsf{X}_{i,o}}{\widehat{\mathrm{E}}_{\cdot,g}} \quad \text{and}$$

$$\widehat{P}_{d,g,o} = \frac{\sum_{i\in\{\}_d\cap\{\}_g} w_i \cdot \mathsf{X}_{i,o}}{\widehat{\mathrm{E}}_{d,g}}$$

for estimation of employment for small domain d: $\mathbf{P}_{d,\cdot,o}$, establishment group g: $\mathbf{P}_{\cdot,g,o}$ and population group g in small domain d: $\mathbf{P}_{d,g,o}$, where

$$\widehat{\mathrm{E}}_{d,\cdot} = \sum_{o=1}^{O} \sum_{i\in\{\}_d} w_i \cdot \mathsf{X}_{i,o}$$

$$\widehat{\mathrm{E}}_{\cdot,g} = \sum_{o=1}^{O} \sum_{i\in\{\}_g} w_i \cdot \mathsf{X}_{i,o} \quad \text{and}$$

$$\widehat{\mathrm{E}}_{d,g} = \sum_{o=1}^{O} \sum_{i\in\{\}_d\cap\{\}_g} w_i \cdot \mathsf{X}_{i,o}$$

are total employment estimates at different levels. Since we have defined the form for $\widehat{\mathsf{X}}^{(3)}_{d,o}$, we substitute the above terms into it and rearrange terms to arrive at exactly the expression

for Estimator 3 in (4), that is,

$$
\begin{aligned}
\widehat{\mathsf{X}}^{(3)}_{d,\,o} &= \widehat{\mathsf{X}}^{(2)}_{d,\,o} + \sum_{g=1}^{G} \widehat{\mathrm{E}}_{d,\,g}(\widehat{P}_{d,\,g,\,o} - \widehat{P}_{\cdot,\,g,\,o}) \\
&= \sum_{g=1}^{G} \mathbf{E}_{d,\,g}\widehat{P}_{\cdot,\,g,\,o} + \sum_{g=1}^{G} \mathbf{E}_{d,\,\cdot}\frac{\widehat{\mathrm{E}}_{d,\,g}}{\widehat{\mathrm{E}}_{d,\,\cdot}}\left(\frac{\widehat{\mathrm{E}}_{\cdot,\,g}\widehat{P}_{\cdot,\,g,\,o}}{\widehat{\mathrm{E}}_{d,\,g}} - \widehat{P}_{\cdot,\,g,\,o}\right) \\
&= \sum_{g=1}^{G} \mathbf{E}_{d,\,g}\widehat{P}_{\cdot,\,g,\,o} + \sum_{g=1}^{G} \frac{\mathbf{E}_{d,\,\cdot}}{\widehat{\mathrm{E}}_{d,\,\cdot}}\left(\widehat{\mathrm{E}}_{\cdot,\,g} - \widehat{\mathrm{E}}_{d,\,g}\right)\widehat{P}_{d,\,g,\,o} \\
&= \sum_{g=1}^{G} \mathbf{E}_{d,\,g}\widehat{P}_{\cdot,\,g,\,o} + \frac{\mathbf{E}_{d,\,\cdot}}{\widehat{\mathrm{E}}_{\cdot,\,\cdot}} \sum_{g=1}^{G} \widehat{\mathrm{E}}_{d,\,g}(\widehat{P}_{d,\,g,\,o} - \widehat{P}_{\cdot,\,g,\,o}) \\
&= \mathbf{E}_{d,\,\cdot} \sum_{g=1}^{G} \frac{\mathbf{E}_{d,\,g}}{\mathbf{E}_{d,\,\cdot}}\widehat{P}_{\cdot,\,g,\,o} + \frac{\mathbf{E}_{d,\,\cdot}}{\widehat{\mathrm{E}}_{d,\,\cdot}} \sum_{g=1}^{G} \widehat{\mathrm{E}}_{d,\,g}\widehat{P}_{d,\,g,\,o} - \frac{\mathbf{E}_{d,\,\cdot}}{\widehat{\mathrm{E}}_{d,\,\cdot}} \sum_{g=1}^{G} \widehat{\mathrm{E}}_{d,\,g}\widehat{P}_{\cdot,\,g,\,o} \\
&= \mathbf{E}_{d,\,\cdot}\left[\sum_{g=1}^{G} \frac{\widehat{\mathrm{E}}_{d,\,g}\widehat{P}_{d,\,g,\,o}}{\widehat{\mathrm{E}}_{d,\,\cdot}} + \sum_{g=1}^{G}\left(\frac{\mathbf{E}_{d,\,g}}{\mathbf{E}_{d,\,\cdot}} - \frac{\widehat{\mathrm{E}}_{d,\,g}}{\widehat{\mathrm{E}}_{d,\,\cdot}}\right)\widehat{P}_{\cdot,\,g,\,o}\right] \\
&= \mathbf{E}_{d,\,\cdot}\left[\widehat{P}_{d,\,\cdot,\,o} + \sum_{g=1}^{G}\left(\frac{\mathbf{E}_{d,\,g}}{\mathbf{E}_{d,\,\cdot}} - \frac{\widehat{\mathrm{E}}_{d,\,g}}{\widehat{\mathrm{E}}_{d,\,\cdot}}\right)\widehat{P}_{\cdot,\,g,\,o}\right]
\end{aligned}
$$

which is exactly in the form of equation (4). This proves equation (4).

References

[1] Cochran, W.G. . *Sampling Techniques*. Wiley, New York, 3rd edition, 1977.

[2] Gonzales, M. E. (1973). *Use and evaluation of synthetic estimators. In Proceedings of the Social Statistics Section*, 33-36. American Statistical Association, Washington, DC.

[3] Li, B. (2001) *Basic Sample Survey Principles & Methods*. SSA Course Note on Survey Sampling Methodology, Sacramento, CA.

[4] NATIONAL CENTER FOR HEALTH STATISTICS (1968). *Synthetic State Estimates of Disability*. P.H.S. Publication 1759. U.s. Government Printing Office, Washington, DC.

[5] Särndal, C.E., E. Swensson and J. Wretman (1992) *Model Assisted Survey Sampling*.

[6] Särndal, C.E. and M. A. Higigrolou (1989), "Small Domain Estimation: A Conditional Analysis," *Journal of the American Statistical Association*, **84**, 266-275.