Wage Imputation in the OES Survey: A Model-Assisted Approach Incorporating Data from the Quarterly Census of Employment and Wages

Jane G. Osburn

Office of Occupational Statistics and Employment Projections
U.S. Bureau of Labor Statistics

1. Introduction

Wage imputation in the BLS Occupational Employment Statistics (OES) Survey requires a process that matches non-respondent and respondent establishments on characteristics that best predict the wage levels of the non-respondent establishments in a given MSA /Industry /Establishment Size cell. The mean wage distribution of the establishments in the donor cell is then imputed to the non-respondent. Currently, the OES Survey procedures first define the donor cell for a given non-respondent establishment /occupation cell by the same time period /MSA / industry (four or five digit NAICS) /establishment-size as the non-respondent, and cells are then collapsed across industry and size groupings in the case of insufficient response. If insufficient response still exists after collapsing across industry and size groupings, the base level strata default from MSA to State, the donor cell is reformed by State / four/five digit NAICS Industry / Establishment Size, and the process repeats as for the MSA base-level strata.[1] The base-level strata eventually default to Nation in a similar manner.

In those instances in which State or Nation are the base-level strata of the donor cell used in an imputation, the current method in effect uses the average wage levels of the State or Nation as a proxy for the wage level of the non-respondent establishment, given industry and establishment size. In the case of states with widely varying area wage levels, the use of the 'State' base level strata can lead to systematically biased individual imputations in many areas.

In the case of the smallest states, the base level strata 'State' is often associated with a donor pool that is only slightly expanded beyond that associated with the 'MSA' base level stratum. In these cases, the number of imputations conducted using the 'Nation' base level stratum is larger than is the case for a large state, and the quality of individual imputation outcomes is diminished in proportion to the divergence between wage levels in the area and the national average.

The problems discussed above do not significantly affect the quality of the wage estimates for most Area /Occupation cells in the OES Survey because MSA is the base level stratum used in the imputations in the vast majority of individual imputations. The

[1] The OES strata include Ownership in NAICs 611, Educational Services, and NAICS 622, Hospitals.

problems affect the quality of the estimates primarily in the case of the small areas located in states that have widely varying area wage levels. Currently in the OES Survey, individual Area /Occupation wage estimates are suppressed if the percentage standard error exceeds a given cutoff. In the case of the smallest areas, only a very small percentage of Area /Occupation estimates pass the criterion, and most estimates are published at the Area / Occupational Group level. A shift to a more accurate method of wage imputation, perhaps used in conjunction with other revisions to the OES Survey design, may significantly expand the domain of area /occupations cells for which high quality estimates are achievable. Alternatively, a shift to a more accurate method of wage imputation may help maintain the quality of the OES wage estimates in the event of changes to the OES Survey design aimed at achieving other estimation targets and goals.

The proposed method of wage imputation replaces the base-level strata currently used in the OES method with successively more aggregated establishment groupings based on statistical estimates of the wage levels of individual establishments. For example, the most disaggregated of these, group set 1, is a thirteen-level grouping in which each group contains establishments with very similar predicted wage levels. Group set 2 is a more aggregated five-level grouping (of all establishments) in which each group contains establishments with wage levels within a broader range of each other. Group set 3 is an even more aggregated, three-level grouping, in which each group contains establishments with wages within an even broader range of each other. Wage imputation follows the same rules used in the current OES procedures, with the exception that group sets one and two take the place of the base-level strata State, Nation in the current OES method, and group set three adds an additional level to the base-level strata.

The focus of the current project is the accurate estimation of area wage levels and of the wage levels of individual establishments in the OES Survey, using the OES data in conjunction with auxiliary data on establishment wages obtained from the Quarterly Census of Employment and Wages Program (QCEW).

Lohr and Prasad (2003, subsequently LP) incorporated information from multiple sources into the estimation of a set of variables of interest by adapting the nested error linear model for the case of multivariate data. The LP estimator is simple to implement and yields large efficiency gains, by effectively ‘borrowing strength’ from variables that are highly correlated with the variable of interest.

The current project uses the LP model to predict area wage levels and the wage differentials of individual respondent and non-respondent establishments in the OES Survey. Simulation is used to examine the performance of wage imputation based on the new approach relative to the current approach and a variety of other alternatives. The results suggest that the proposed approach provides a significant expansion in the number of areas for which high quality estimates of occupational wages can be obtained.

The following section discusses the data produced by the Occupational Employment Statistics Survey. Section three outlines the prediction of area wage levels using the Lohr-Prasad model, as well as the prediction of individual establishment wage levels. Section four describes the simulation. Section five discusses the performance measures used to evaluate the alternative models. Section six discusses the results, and section seven offers some conclusions.

2. Data from the Occupational Employment Statistics Survey

The Bureau of Labor Statistics Occupational Employment Statistics Survey collects data over a three year cycle on the wages and occupational category of each employee in a total of approximately 1.2 million establishments, spanning the non-farm private and public sectors in the U.S. For each occupation employed in the establishment, survey respondents record the number of employees earning wage rates that fall within each of twelve wage intervals. The average wage of an Establishment /Occupation cell is the sum over wage intervals of the product of the employment in the interval and the 'midpoint', or estimated mean wage of the interval, divided by the sum of employment across the wage intervals. The wage interval midpoints that are applied to establishments surveyed during different quarters of the three year period are updated using the Employment Cost Index (ECI), such that all establishments' wages reflect the final year /quarter wage levels in the completed dataset.[2]

3. Multivariate EBLUP Estimation

3.1 The Lohr-Prasad Estimator

The main component of multivariate EBLUP estimation that sets it apart from univariate EBLUP estimation is the role of covariance components. In the current context of a two-level multilevel model of wages in which the level two units consist of areas and the level one units consist of establishment /occupation wages, covariance components refer to the covariance within areas of wages obtained from two or more separate surveys. In cases where these data are available for matched establishments, estimates of components of covariance can be obtained at both levels one and two, and the multivariate estimator borrows strength from the auxiliary data at both levels.

A multivariate EBLUP of the area effect of wages is similar in structure to a univariate EBLUP estimator; it is a weighted average residual wage for the area that is shrunk toward zero by an amount proportional to the contribution of areas (i.e. of model variance) to the explanation of total wage variance. The multivariate 'shrinkage'

[2] The midpoints of the wage intervals are estimated using data from the Bureau of Labor Statistics National Compensation Survey.

parameter $\sum_{j=1}^{n_i} w_{ij}\left(\Sigma_{\upsilon}^{-1}+\mathrm{E}_i\right)^{-1}\Sigma_e^{-1}$ is similarly composed of the multivariate variance components Σ_{υ} and Σ_e.

where

$$\Sigma_{\upsilon}=\begin{bmatrix}\Sigma_{\upsilon oo} & \Sigma_{\upsilon oq}\\ \Sigma_{\upsilon qo} & \Sigma_{\upsilon qq}\end{bmatrix},\quad \Sigma_e=\begin{bmatrix}\Sigma_{eoo} & \Sigma_{eoq}\\ \Sigma_{eqo} & \Sigma_{eqq}\end{bmatrix}$$

$\Sigma_{\upsilon oo}$ = the covariance within areas of OES establishment residual wages (log wage minus the fixed effects of occupation, industry, and establishment size)

$\Sigma_{\upsilon qq}$ = the covariance within areas of QCEW establishment residual wages

$\Sigma_{\upsilon qo}$ = the covariance within areas of QCEW and OES establishment residual wages

Σ_{eoo} = the variance within areas of OES conditional establishment wage residuals, obtained by subtracting the area effect of wages from the OES residual wage

Σ_{eoq} = the covariance within establishments of the QCEW and OES conditional wage residuals

Σ_{eqq} = the variance within areas of the QCEW conditional establishment wage residuals

$$\mathrm{E}_i = n_i\,\Sigma_e^{-1}+n_i^*\left(\Sigma_e^*\right)^{-1}$$

The Lohr-Prasad estimator is;

1) $$\hat{\tilde{\mu}}_i=\mu+\sum_{j=1}^{n_i} w_{ij}\left(\Sigma_{\upsilon}^{-1}+\mathrm{E}_i\right)^{-1}\Sigma_e^{-1}\left(\bar{u}_i-\mu\right)_{XY}+\left(\Sigma_{\upsilon}^{-1}+\mathrm{E}_i\right)^{-1}n_i^*\left(\Sigma_e^*\right)^{-1}\left(\bar{u}_i^*-\mu\right)_Y$$

where

$$\left(\bar{u}_i-\mu\right)_{XY}=\begin{bmatrix}\bar{u}_{i\,o}-\mu_o\\ \bar{u}_{i\,q}-\mu_q\end{bmatrix}_{XY}\qquad \left(\bar{u}_i^*-\mu\right)=\left[\bar{u}^{*}{}_{i\,q}-\mu_q\right]_{XX}$$

$$\dot{E}d_{j\,OES}=\left[\ln\left(mean\,OES\ estab\,hourly\ wage\right)-\mathrm{int}\,ercept-\left(ind\hat{u}stry\right)-\left(si\hat{z}e\right)\right]$$

$\bar{u}_{i\,o}-\mu_o$ = the area mean of the residuals $\dot{E}d_{j\,OES}$;

$$\dot{E}d_{j\,naive\_QCEW}=\left[\ln\left(mean\,QCEW\ estab\,hourly\ wage\right)-\mathrm{int}\,ercept-\left(ind\hat{u}stry\right)-\left(si\hat{z}e\right)\right]$$

where the mean QCEW estab hourly wage = total quarterly wages divided by total quarterly employment, coverted to an hourly wage using a (naivc) assumed forty hour work week for all employees.

$\bar{u}_{i\,q}-\mu_q$ = the area mean of the QCEW wage residuals $\dot{E}d_{j\,naive\_QCEW}$

$\left[\bar{u}^{*}_{iq} - \mu_q\right]_{XX}$ = the area mean of $\dot{E}d_{j\ naive\_QCEW}$ in the data subset containing establishments for which only QCEW data exists.

The subscript XY indicates that the variable is constructed using only those establishment observations for which both the QCEW and OES wages exist. Similarly, the subscript XX indicates that the variable is constructed using only those establishment observations for which only the QCEW wages exist.

The residuals $\dot{E}d_{jnaive\_QCEW}$ and $\dot{E}d_{j\ OES}$ with means

$$\left(\bar{u}_i - \mu\right)_{XY} = \begin{bmatrix} \bar{u}_{io} - \mu_o \\ \bar{u}_{iq} - \mu_q \end{bmatrix}_{XY} \quad \text{and} \quad \left(\bar{u}^{*}_i - \mu\right) = \left[\bar{u}^{*}_{iq} - \mu_q\right]_{XX}$$

are often termed *multilevel residuals*, because they are usefully conceived of as containing the residuals from both levels 1 and 2 of the multilevel model. In other words, they consist, for each observation, of the area wage effect plus the residual error for the individual establishment.

The following sections explain how the OES and QCEW data are used to obtain multivariate EBLUP prediction estimates of the area wage effect and of the 'conditional' portion of the establishment wage differential, i.e. the deviation of the establishment average wage from that of the area.

3.2. Robust Estimation of Multivariate Variance Components

The main task of constructing the Lohr-Prasad estimator is estimating the variance /covariance components. These are then used together with the area mean multilevel residual wage to assemble the estimate. Dueck and Lohr (2005) warn that estimates of covariance components are highly sensitive to outliers, and should be estimated using robust methods. In general, the focus of robust methods is to correct the outlier problems in the original estimation dataset using information about the individual observations together with information about their variances /covariances. The resulting 'pseudo' dataset is then used to estimate the variable of interest.

First, a combined OES-QCEW establishment-level dataset is constructed using the establishment respondents in the May 2011 Survey round. This dataset contains matched OES / QCEW data for establishments from each of the six panels up to and including the May 2011 panel. In the QCEW portion of this dataset, the analysis variable is a naïve mean establishment wage constructed as total quarterly wages, scaled to represent total hourly wages, divided by the total number of employees. As such, this naïve mean wage does not control for worker hours or occupation. Each establishment-level observation is also identified by industry, area, year, etc.etc. Creation of the OES portion of the dataset

required a preliminary analysis to create establishment-level observations from the existing Establishment /Occupation level observations.

The wage data in the combined dataset are adjusted to reflect wage levels in the current reference period. The OES Survey 'updates' the wages in the five earlier panels to reflect wage levels in the current reference period using the Employment Cost Index. The QCEW wage data from earlier panels were updated in a similar manner, using establishment-level update factors constructed as an employment-weighted average of the State /Occupation level update factors used in the OES Survey.

The combined OES-QCEW establishment-level dataset is used to obtain preliminary estimates of the variance components. In this estimation, the multivariate dependent variable 'lnwage $_{\text{wage type}}$' contains, in separate observations, each of the two wages discussed above;

1) the mean log OES establishment wage
2) the mean log QCEW 'naive' establishment wage

The estimation flexibly allows for differences across these two dependent variables in the effects of each predictor ;

1) $lnwage_{ij,type} = intercept + industry + establishment\ size_j + (Intercept * type) + (industry * type) + (establishment\ size_j * type) + (area * type) + e_{ijtype}$

where
i indexes areas, j indexes establishments, type is an indicator for wage type

Area $_{\text{wage type}}$ ~N(0 0, Σ_{υ})

e $_{\text{wage type}}$ ~N(0 0, Σ_e)

Due to computer resource constraints, an unweighted analysis was used to estimate the variance components. Pfeffermann et. al. (1998) have shown that estimates of the variance components are consistent in this case as long as all of the survey strata variables are included as covariates. The output from this analysis are preliminary estimates of the variance/covariance components and the set of multilevel residuals discussed earlier.

The multilevel residuals obtained from this analysis are first standardized using a spectral decomposition of the variance/covariance component estimates $\hat{\Sigma}_{\upsilon}$ and $\hat{\Sigma}_{\mathrm{e}}$. Spectral decomposition provides a convenient means of standardizing residuals by factorizing the covariance matrix into scalar (eigenvalue) and vector (eigenvector) components. The general form of these standardized residuals for an establishment j in area i is;

$r_{ij} = \left[\hat{\lambda}^{-1/2}\, C'(y_{ij} - \hat{\mu}_{ij})\right]$ where the establishment-level multilevel residual is here denoted by $(y_{ij} - \hat{\mu}_{ij})$, to distinguish it from the mean multilevel residual mentioned earlier.

$\hat{\lambda}$ = the diagonal matrix of eigenvalues of the matrix Σ_{υ} or Σ_{e} and

C = the associated eigenvectors of the matrix Σ_{υ} or Σ_{e}

Next, the standardized residuals are winsorized, and the standardized, winsorized residuals are used to construct robust estimates $\hat{u}_i$, of $\overline{u}_i$ (the area mean wage) and $\hat{u}$ of u (the overall mean wage) similar to those that appear in 1) on page 4. Note, however, that the fixed effects estimates from the un-weighted analysis of 1) are not consistent, and the estimates $\hat{u}_i$, $\hat{u}$ are not consistent either; they are used solely for the purpose of obtaining robust estimates of the variance components.

M-estimation is used to obtain the estimates $\hat{u}_i$, $\hat{u}$ from the transformed residuals. This estimation avoids the use of maximum likelihood, which requires full specification of the distribution of the random variable under consideration and is inconsistent in the case of misspecification of this distribution. For the estimation of $\overline{u}_i$, the multivariate multilevel residuals are standardized using a spectral decomposition of $\hat{\Sigma}_e$, and the estimate $\hat{u}_i$ is the solution to the implicit function;

$$\sum_{j=1}^{n_i} \Psi\left[\hat{\lambda}^{-1/2} C'(y_{ij} - \hat{u}_i)\right] = 0$$

$$\text{where } \Psi = \begin{cases} x & if\ |x| \le c \\ sign(x) * c & else \end{cases}$$

n_i = the number of establishments contributing to the estimate for area i

For the estimation of $\hat{u}$, the multivariate residuals are standardized using a spectral decomposition of $\hat{\Sigma}_{\alpha}$, and the estimate $\hat{u}$ is the solution to the implicit function;

$$\sum_{i=1}^{M} n_i \Psi\left[\hat{\lambda}_{\propto}^{-1/2} C'_{\propto}(\hat{u}_i - \hat{u})\right] = 0$$

n_i = the number of establishments contributing to the estimate for area i

The estimates $\hat{u}_i$ and $\hat{u}$ are then used to get $\propto_i = \hat{u}_i - \hat{u}$.

The aforementioned components are then used to assemble pseudo observations;

$$\hat{y}_{ij} = \hat{u} + \hat{C}_{\propto} \hat{\lambda}_{\propto}^{-1/2} K_{\propto} \psi \left[\hat{\lambda}_{\propto}^{-1/2} \hat{C}_{\propto}{}' \, \hat{\propto}_{\iota} \right] + \widehat{\hat{C} \hat{\lambda}^{-1/2}} K_e \psi \left[\hat{\lambda}^{-1/2} \hat{C}' (y_{\iota J} - \hat{u}_i) \right]$$

The pseudo observations are then used to estimate the variance components, using REML estimation, in an analysis similar to 1).

3.3. Assembling the Lohr-Prasad Estimator and Obtaining Predictions of the Conditional Establishment Wage Differential

Next, the survey-weighted area means of the full set of multilevel residuals $\left(\overline{u}_i - \mu\right)_{XY}$ and $\left(\overline{u}_i{}^{*} - \mu\right)_{XX}$ are obtained using an analysis similar to 1), except the analysis is weighted and there is no modeling of the residual variance.[3] The robust estimates of the variance components $\hat{\Sigma}_{\upsilon}$ and $\hat{\Sigma}_{e}$ are used together with these area mean residuals to assemble the EBLUPs of the OES and QCEW wage effect of area i, $\hat{\mu}_i$ in 1) .

1`) $\hat{\mu}_i = \left(\Sigma_{\upsilon}^{-1} + \mathrm{E}_i\right)^{-1} \Sigma_e^{-1} \left(\overline{u}_i - \mu\right)_{XY} + \left(\Sigma_{\upsilon}^{-1} + \mathrm{E}_i\right)^{-1} n_i^{*} \left(\Sigma_e^{*}\right)^{-1} \left(\overline{u}_i^{*} - \mu\right)_{XX}$

The multivariate EBLUP of the OES area wage effect is the sum of three terms; a univariate shrinkage factor multiplied by the mean OES multilevel residual of the matched establishments, a univariate shrinkage factor (multiplied by -1) multiplied by the mean QCEW multilevel residual of the matched establishments, and a univariate shrinkage factor multiplied by the mean QCEW multi-level residual of the non-matched establishments. The multivariate EBLUP of the QCEW area wage effect is obtained from 1) in a similar manner.

For the purpose of predicting the wage differential of individual establishments, it is useful to think of the wage differential as composed of the sum of the area wage level and the 'conditional' establishment wage differential, where the latter is just the difference between the wage differential of the establishment and the area wage effect.

The prediction of the conditional establishment wage differential is a QCEW-based projection estimate constructed using the residual variance/ covariance parameters that were estimated in 1);

$$\hat{E}d_{\_OES} = \hat{\mu}_{io} + \beta_{XY,estab}(Ed_{i\,j\,\_naive\ QCEW})$$

where $\beta_{XY,estab} = \dfrac{\Sigma_{eoq}}{\Sigma_{eqq}}$

[3] The analysis would not run with both weighting and modeling of the residual variance.

4. The Simulation

Simulation is used to test the various model alternatives against the OES method. The main elements of the simulation include a 'master'file that contains, for each alternative wage model, identifiers for the establishment groupings one, two, and three that apply for each establishment in the OES dataset. That is, for each wage model separately, each establishment is classified into one of the 13 levels of group1, one of the five levels of group 2, and one of the three levels of group 3.

As was discussed in Section 3.3., the model prediction of the establishment wage differential is composed of an estimate of the area effect of wages and an estimate of the 'conditional' establishment wage differential. The first model, model G, consists of the use of the OES multivariate EBLUP to predict the area effect of wages and the use of the model discussed at the end of section 3.3. to predict the conditional establishment wage differential. This model is used to predict the wage differential of each establishment in the OES dataset. The information that is placed on the master file contains the groups one, two, and three classifications from this model for every establishment in the OES dataset.

Model UNI uses the OES univariate EBLUP as the prediction of the establishment wage differential. Model UNI-Q uses the univariate EBLUP to predict the area effect of wages and the model discussed at the end of section 3.3. to predict the conditional establishment wage differential. Model LP uses the OES multivariate EBLUP as the prediction of the establishment wage differential. Model LP-Q uses the multivariate EBLUP to predict the area effect of wages and the model discussed at the end of section 3.3. to predict the conditional establishment wage differential.

Forty simulated imputations of the May 2011 OES dataset were conducted using those establishments that could be matched to the QCEW dataset (the majority of establishments). The simulated OES survey sample is comprised of all of the respondent establishments in the dataset. In each simulated imputation, a stratified random sample is taken from the simulated sample and used to identify a set of simulated non-respondent establishments that makeup a similar proportion of establishments and a similar Industry /Size composition of establishments to the set of actual non-respondent establishments in the actual OES sample. The simulated respondents include all units in the simulated sample not identified as simulated non-respondents.

The simulation for each model is guided by the 'Imputation dataset', which contains the simulated non-respondent establishment data stripped of all wage distributions, but retaining identifiers including a non-respondent indicator, industry and establishment size identifiers, establishment total employment, the occupational employment of each occupation employed in the establishment, and the group one, two, and three assignments under each of the models.

OES procedures first identify respondent establishments for a given semi-annual panel and occupation from the same MSA/ four/five digit NAICS Industry /Size(2) cell as the non-respondent establishment. If the cell contains a sufficient number of respondents, the mean wage distribution of the cell is imputed to the non-respondent Establishment /Occupation cell. In the absence of sufficient response, the cells are combined into successively more aggregated industry and size groupings, and the mean wage distribution of the aggregated grouping is used. If cells defined by MSA /All Industries /Occupation still contain insufficient response, the base-level strata defaults from MSA to State, the cell is reformed by State /four/five digit NAICS Industry / Size (2) /Occupation, and the procedure repeats as for the MSA base-level strata. If cells defined by State /All Industries /Occupation still contain insufficient response, the base-level strata defaults to USA, cells are reformed by USA /four/five digit NAICS Industry /Size (2) /Occupation, and the procedure repeats as for the State base-level strata.

Imputation under each of the alternative models follows parallel procedures. The only difference is that the base level strata used in the OES (i.e. MSA /State /USA) are replaced with base level strata (MSA /group set 1 / group set 2 / group set 3) in each of the alternative models.[4] The mean wage distribution of the donor cell is then used to apportion the recipient cell occupational employment among the OES wage intervals.[5]

For each simulation and for each of the six experimental models and the OES method, the mean imputed wage by Establishment /Occupation is calculated using the wage interval means in conjunction with the apportioned occupational employment.

5. Performance Measures

Performance measures including average relative error (ARE) average relative bias $\overline{(ARB)}$, and average relative efficiency (RLEF) are calculated. Due to the small number of simulations, the quality of the performance estimates relies crucially on the large number of Area /Establishment /Occupation cells that is imputed in each simulation.

Let

$\overline{w}_{sim}$ = mean imputed wage by Sim /MSA /Establishment /Occupation

$\overline{w}_{actual}$ = mean "true" wage by MSA /Estab. /Occ. (wage of the respondent estab/occ cell) in the actual respondent establishment.

[4] All simulated imputations are conducted by semi-annual panel, as in the current procedures.

[5] In the actual OES procedures, non-respondent establishments are first employment-imputed and then wage imputed. The simulation by-passes the employment-imputation step in favor of using the actual occupational employment totals of the simulated non-respondents that are available by virtue of the simulation design.

nocc = number MSA /Establishment /Occupation cells in summary cell
t = un-weighted occupational employment by MSA /Establishment /Occupation

$$ARE = \left(\frac{1}{\sum_{1}^{nocc}(t)}\right)\sum_{1}^{nocc}\left(\frac{1}{40}\right)\sum_{sim=1}^{40}\left(t * \left|\frac{\overline{w}_{sim}}{\overline{w}_{actual}} - 1\right|_{OCC}\right)$$

$$ARB = \left(\frac{1}{\sum_{1}^{nocc}(t)}\right)\sum_{1}^{nocc}\left(\frac{1}{40}\right)\sum_{sim=1}^{40}\left[t * \left(\frac{\overline{w}_{sim}}{\overline{w}_{actual}} - 1\right)\right]$$

$$\overline{RLEF} = MSE\ (Experimental\ \ Estimator\ ) / MSE\ (OES\ Estimator\ )$$

$$MSE\ (Estimator) = \left(\frac{1}{\sum_{1}^{c=number\ Area/occ\ \ cells}(t)_c}\right)\sum_{1}^{c}\left(\frac{1}{40}\right)\sum_{sim=1}^{40}\left[t * \left(est_{c,sim} - Y_c\right)^2\right]$$

Y_c is the actual mean (reported) wage for the cell.

6. Results

Table 1 provides measures of average relative error, average relative bias, and average relative efficiency overall and by area size. The results show that the experimental estimators improve on the OES estimator by the greatest amounts in the unweighted versions of the measures, which effectively weight each establishment /occupation cell equally, suggesting that the performance gains of the experimental estimators are concentrated in the smaller areas and occupations. The proposed estimator (LP_Q) has about nine percent lower unweighted average relative error than the OES estimator, and between thirty and seventy five percent lower unweighted average relative bias, with the steepest gains in the smallest areas. The weighted versions of these estimators show only a slight improvement over the OES estimator for most areas, and a significant deterioration in performance for the smallest areas.

Table 2 provides summary measures of average relative error, average relative bias, and average relative efficiency by Establishment size. The average relative bias of the proposed estimator is nearly fifty percent below that of the OES estimator for the smallest establishments in both the weighted and unweighted versions of the measures. These bias gains all but disappear as the establishments get larger, until the bias of the proposed estimator actually exceeds that of the OES for the largest establishments, in both the weighted and unweighted versions of the measures. The relative efficiency measures show that the proposed (LP-Q) estimator has smallest MSE relative to the OES estimator for the smallest establishments and is always smaller than that of the OES estimator.

Table 1. Results Overall and by Area Size

Area	Emp. Weighted Average Relative Bias					Unwtd Average Relative Bias				
Size	UNI	UNI_Q	LP	LP_Q	OES	UNI	UNI_Q	LP	LP_Q	OES
ALL	0.09	0.07	0.06	0.09	0.09	0.07	0.07	0.06	0.06	0.09
< 100K	0.05	0.07	0.05	0.08	0.11	0.07	0.07	0.07	0.06	0.13
<50K	0.04	0.04	0.03	0.06	0.06	0.09	0.05	0.09	0.05	0.11
<25K	-0.09	-0.17	-0.09	-0.12	0.03	0.09	0.00	0.12	0.03	0.13

Area	Emp.Weighted Average Relative Error					Unwtd Average Relative Error				
Size	UNI	UNI_Q	LP	LP_Q	OES	UNI	UNI_Q	LP	LP_Q	OES
ALL	0.31	0.28	0.30	0.27	0.30	0.31	0.31	0.30	0.30	0.33
< 100K	0.21	0.22	0.21	0.22	0.22	0.27	0.27	0.27	0.26	0.29
<50K	0.20	0.19	0.20	0.19	0.18	0.30	0.25	0.30	0.24	0.29
<25K	0.22	0.27	0.22	0.22	0.18	0.29	0.22	0.30	0.23	0.29

Area	Efficiency Relative to OES Metho			
Size	UNI	UNI_Q	LP	LP_Q
ALL	1.00	0.97	1.03	0.94
< 100K	0.96	0.98	1.01	0.97
<50K	1.01	0.95	1.04	0.91
<25K	1.10	1.31	1.17	1.13

Table 2. Results by Establishment Size

Estab	Emp. Weighted Average Relative Bias					Unwtd Average Relative Bias				
Size	UNI	UNI_Q	LP	LP_Q	OES	UNI	UNI_Q	LP	LP_Q	OES
1	.079	.099	.097	.076	.135	.082	.114	.111	.079	.158
2	.077	.069	.070	.077	.086	.074	.066	.066	.074	.090
3	.078	.095	.091	.079	.097	.077	.090	.088	.080	.107
4	.069	.076	.075	.070	.083	.067	.072	.070	.069	.093
5	.050	.039	.037	.048	.033	.045	.031	.030	.044	.053

Estab	Emp.Weighted Average Relative Error					Unwtd Average Relative Error				
Size	UNI	UNI_Q	LP	LP_Q	OES	UNI	UNI_Q	LP	LP_Q	OES
1	.278	.307	.306	.277	.329	.327	.363	.361	.326	.388
2	.249	.263	.264	.249	.277	.296	.307	.307	.297	.324
3	.237	.257	.255	.239	.266	.288	.299	.299	.289	.316
4	.234	.246	.246	.235	.255	.286	.291	.291	.286	.308
5	.245	.242	.243	.242	.251	.279	.274	.275	.279	.293

Estab	Relative Efficiency			
Size	UNI	UNI_Q	LP	LP_Q
1	.903	.972	.970	.902
2	.927	.970	.970	.932
3	.931	.974	.958	.936
4	.920	.971	.965	.918
5	.984	.955	.966	.977

Size	Estab. Emp	
1	< 10	
2	>9	<20
3	>19	<50
4	>49	<250
5	>249	

7. Discussion and Conclusions

The proposed method of wage imputation significantly increases the number of small and very small areas for which high quality wage imputations can be accomplished in the OES Survey. These gains were similar for the UNI-Q (univariate ELBUP) and the LP-Q (multivariate EBLUP) models, with the LP-Q estimator yielding a slight improvement in relative efficiency over the UNI-Q model.

In the case of the smallest establishments, the relatively large performance gains of the 'Q' estimators are the result of relatively high correlation between the naïve establishment wage differential obtained from the QCEW average establishment wage and the more accurate establishment wage differential that is constructed using the OES data and that controls for occupation. This correlation weakens with establishment size until, for the largest establishments, the relatively high variance of the QCEW component dominates the information gains. Table 2 shows that, for the largest establishments, the weighted average relative bias of the LP-Q estimator actually exceeds that of both the LP and the OES estimators,

The approach examined here is most useful for the purpose of imputing wages for small and very small areas, but also holds the potential to help maintain estimate quality in the case of OES funding disruptions. In the case of a reduction in the OES sample size, the methods examined here could gain importance with respect to areas larger than those for which the LP-Q model performed best in this simulation. In the case of such "effectively small" areas, use of the multivariate approach examined here can help maintain estimate quality. On the other hand, if future changes in the OES Survey bring a renewed focus on small area estimation, the approach examined here offers a means of significantly expanding the number of domains considered.

References

Abayomi, Kobi, Andrew Gelman, and Marc Levy 2008. "Diagnostics for Multivariate Imputations," Applied Statistics, Vol.57., Part.3. pp. 273-271.

Asparouhov, Tihomir 2005. "Sampling Weights in Latent Variable Modeling," *Structural Equation Modeling*, Vol. 12., No.3., pp.411-434.

Datta, G.S., B. Day, and I. Basawa 1999. "Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation," *Journal of Statistical Planning and Inference*, 75, pp.269-279.

Datta, G.S. and P. Lahiri 2000. "A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems," *Statistica Sinica*, 10, pp.613-627.

Deuck, A. and S. Lohr 2005. "Robust Estimation of Multivariate Covariance Components," *Biometrics*, 61, pp.162—169.

Eltinge, John L. and Jeffrey M. Gonzalez 2009. "Sensitivity of Inference Under Imputation," Federal Committee on Statistical Methodology Research Conference, Washington, DC.

Goldstein, Harvey 2011. *Multilevel Statistical Models*, 4th Edition, John Wiley and Sons, Ltd.,
Chichester, West Sussex, UK.

Greene, William H. 2000. *Econometric Analysis*, Fourth Edition. Prentice Hall Press, New Jersey.

Grilli,Leonardo and Monica Pratesi. 2002. "Weighted Estimation in Multilevel Models to Allow for Informativeness of the Sampling Design," *Working Paper 2002 /01*, Dipartimento di Statistica, Università degli Studi di Firenze.

Hox, Joop. 2002. *Multilevel Analysis: Techniques and Applications*, Lawrence Erlbaum Associates, Mahwah, NJ.

Littell, R.C., George Milliken, Walter Stroup, Russell Wolfinger, and Oliver Schabenberger 2006. *SAS for Mixed Models*, Second Edition, SAS Institute Inc., Cary, North Carolina.

Lohr, S and Prasad, N.G.N. 2003. "Small Area Estimation With Auxiliary Survey Data," *Canadian Journal of Statistics*, Vol.31, No.4., pp. 383-396.

Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein, J. Rasbach "Weighting for Unequal Selection Probability in Multilevel Models," *Journal of the Royal Statistical Society*, B, 1998 Vol.60., Part 1. pp.23-40.

Prasad, N.G.N and J.N.K.Rao. 1990. "The Estimation of Mean Square Error of Small Area Estimators," *Journal of the American Statistical Association*, Vol.85., pp.163-171.

Rao, J.N.K. 2003. *Small Area Estimation,* John Wiley and Sons, New Jersey

Royal, R.M. 1976. "The Linear Least-Squares Prediction Approach to Two-Stage Sampling," *Journal of the American Statistical Association*, Vol.71.,pp.657-664.
Shoemaker, Owen and William Johnson. 1999. "Estimation of Variance Components for the U.S. Consumer Price Index," *Proceedings of the Section on Business and Economic Statistics, American Statistical Association.*

Snijders, Tom and Roel Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling,* Sage Publications, Thousand Oaks, CA.

Stukel, D.M. and Rao, J.N.K. 1999. "Small Area Estimation Under Two-fold Nested Error Regression Models," *Journal of Statistical Planning and Inference*, Vol.78., pp. 131-147.

Endnotes

[1] See Pfeffermann, et.al (1998). Weighting is not needed in a linear mixed model that includes as covariates each of the strata components of a self-weighting survey, in which the probability of selection is the same for all units within a stratum. The parameter estimates of the linear mixed model are biased if units at any level (i.e. individual establishment/occupation observations at level 1, MSAs at level 2) are selected with unequal probabilities in ways that are not accounted for by the linear mixed model. While the OES Survey is self-weighting within MSA/Industry/Size strata, the model used here includes these strata components as covariates only at somewhat aggregated levels. For example, a three digit Naics industry is included rather than the four or five digit Naics industry used in sampling.

A correction offered by Pfeffermann, et.al (1998) involves scaling the level 1 sample weights such that the naïve estimate of MSA employment, $\hat{N}_i$, equals actual total reported employment for the area. With this scaling, the variance of the cluster-level latent variable is determined by the true sample size of level 1 units (see Asparouhov (2005). Goldstein (2011) also provides a discussion of the approach. Grilli and Pratesi (2002) conducted extensive simulations demonstrating that the approach outlined above works well, minimizing both bias and the increase in variance that results from the use of survey weights.