Survey Methods and Reliability Statement for the May 2022 Occupational Employment and Wage Statistics Survey

Introduction

The Occupational Employment and Wage Statistics (OEWS) survey measures occupational employment and wage rates for wage and salary workers in nonfarm establishments nationally, and in the 50 states and the District of Columbia, Guam, Puerto Rico, and the Virgin Islands.

About 8.3 million in-scope establishments are stratified within their respective states by substate area, industry, size, and ownership. Substate areas include all officially defined metropolitan areas and one or more nonmetropolitan areas. The North American Industry Classification System (NAICS) is used to stratify establishments by industry.

Probability sample panels of about 179,000 to 187,000 establishments are selected semiannually. Responses are obtained by Internet or other electronic means, mail, email, fax, telephone, or personal visit. Respondents report their employees' job titles and descriptions, which are used to classify workers into occupations in the Standard Occupational Classification (SOC) system. OEWS receives individual wage rate data for the federal government; the U.S. Postal Service; most state governments; and, as of the May 2020 panel, most private sector and local government establishments. For the remaining establishments, the OEWS survey data are placed into 12 wage intervals.

Estimates of occupational employment and wage rates are based on six panels of survey data collected over a 3-year cycle. The final in-scope sample size when six panels are combined is approximately 1.1 million establishments. Total 6-panel unweighted employment covers approximately 80 million out of the total employment of almost 139 million.

The estimates described here are produced using a model-based estimation method using 3 years of OEWS data (MB3) to estimate current year occupational employment and wages. The MB3 estimation method was introduced with the May 2021 OEWS estimates; additional changes to the MB3 wage processing methodology were made for the May 2022 estimates and are discussed below. MB3 estimation has advantages over the previous OEWS methodology, as described in the *Monthly Labor Review* article

Model-Based Estimates for the Occupational Employment Statistics program. The sampling methods described here are the same as those used prior to May 2021.

Occupational and industrial classification systems

The occupational classification system

The U.S. Office of Management and Budget's Standard Occupational Classification (SOC) system is used to define occupations. All panels in the 2022 estimates were collected using the 2018 SOC system. More information about the 2018 SOC system can be found at www.bls.gov/soc/2018/home.htm.

The industrial classification system

The OEWS survey classifies establishments into industries based on the North American Industry Classification System (NAICS). The May 2022 OEWS estimates use the 2022 North American Industry Classification System (NAICS). More information about NAICS can be found at www.bls.gov/bls/naics.htm or in the 2022 North American Industry Classification System manual available at www.census.gov/naics/. Each establishment in the survey is assigned a 6-digit NAICS code based on its primary economic activity.

The May 2022 estimates are the first to be produced using the 2022 NAICS, which replaces the 2017 NAICS used for the May 2017 – May 2021 estimates. All six survey panels used for the May 2022 estimates were collected using the 2017 NAICS codes; these data were then mapped to the corresponding 2022 NAICS codes.

Industrial scope and stratification

The survey covers the following NAICS industry sectors:

11	Logging (1133), support activities for crop production (1151), and support activities for animal production (1152) *only*
21	Mining, quarrying, and oil and gas extraction
22	Utilities
23	Construction

31-33	Manufacturing
42	Wholesale trade
44-45	Retail trade
48-49	Transportation and warehousing
51	Information
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific, and technical services
55	Management of companies and enterprises
56	Administrative and support and waste management and remediation services
61	Educational services
62	Healthcare and social assistance
71	Arts, entertainment, and recreation
72	Accommodation and food services
81	Other services, except public administration [private households (814) are excluded]

Federal government executive branch (assigned industry code 999100)*

State government, excluding schools and hospitals (assigned industry code 999200)*

Local government, excluding schools, hospitals, gambling establishments, and casino hotels (assigned industry code 999300)*

* These are OEWS-defined industry codes and not a part of the NAICS industry classification.

These sectors are stratified into about 300 industry groups at the 3-, 4-, 5-, or 6-digit NAICS level of detail.

Concepts

An ***establishment*** is generally a single physical location at which economic activity occurs (e.g., store, factory, restaurant, etc.). Each establishment is assigned a 6-digit NAICS code. When a single physical location encompasses two or more distinct economic activities, it is treated as two or more separate establishments if separate payroll records are available and certain other criteria are met.

Employment refers to the number of workers who can be classified as full- or part-time employees, including workers on paid vacations or other types of paid leave; salaried officers, executives, and staff members of incorporated firms; employees temporarily assigned to other units; and noncontract employees for whom the reporting unit is their permanent duty station, regardless of whether that unit prepares their paychecks.

The OEWS survey includes all full- and part-time wage and salary workers in nonfarm industries. Self-employed workers, owners and partners in unincorporated firms, household workers, and unpaid family workers are excluded.

Occupations are classified based on work performed and on required skills. Employees are assigned to an occupation based on the work they perform and not on their education or training. For example, an employee trained as an engineer but working as a drafter is reported as a drafter. Employees who perform the duties of two or more occupations are reported in the occupation that requires the highest level of skill or in the occupation where the most time is spent if there is no measurable difference in skill requirements. ***Working supervisors*** (those spending 20 percent or more of their time doing work similar to that of the workers they supervise) are classified with the workers they supervise. ***Workers receiving on-the-job training, apprentices, and trainees*** are classified in the occupations for which they are being trained.

A ***wage*** is money that is paid or received for work or services performed in a specified period. Base rate pay, cost-of-living allowances, guaranteed pay, hazardous-duty pay, incentive pay such as commissions and production bonuses, and tips are included in a wage. Back pay, jury duty pay, overtime pay, severance pay, shift differentials, nonproduction bonuses, employer costs for supplementary benefits, and tuition reimbursements are excluded. The federal executive branch, the U.S. Postal Service (USPS), and

most state governments report individual wage rates as dollars per hour or per year for workers. For the November 2019 panel, most wage data were processed using the 12 wage intervals below in Table 1. Starting in May 2020, individual wage rates were collected from all employers, when available. Employers who did not report individual wage rates could still report wage interval data.

Table 1: Wage Intervals for the November 2019 – May 2022 Survey Panels

Interval	Wages	
	Hourly	Annual
Range A	Under $9.25	Under $19,240
Range B	$9.25 to $11.99	$19,240 to $24,959
Range C	$12.00 to $15.49	$24,960 to $32,239
Range D	$15.50 to $19.74	$32,240 to $41,079
Range E	$19.75 to $25.49	$41,080 to $53,039
Range F	$25.50 to $32.74	$53,040 to $68,119
Range G	$32.75 to $41.99	$68,120 to $87,359
Range H	$42.00 to $53.99	$87,360 to $112,319
Range I	$54.00 to $69.49	$112,320 to $144,559
Range J	$69.50 to $89.49	$144,560 to $186,159
Range K	$89.50 to $114.99	$186,160 to $239,199
Range L	$115.00 and over	$239,200 and over

Three-year survey cycle of data collection

The survey is based on a probability sample drawn from the universe of in-scope establishments stratified by geography, industry, size, and ownership. The sample is designed to represent all nonfarm establishments in the United States.

The OEWS survey allocates and selects a sample of approximately 179,000 to 187,000 establishments semiannually. Semiannual samples are referred to as *panels*. To the extent possible, private sector units selected in any one panel are not sampled again in the next five panels.

The survey is conducted over a rolling 6-panel (or 3-year) cycle. This is done in order to provide adequate geographic, industrial, and occupational coverage. Over the course of a 6-panel (or 3-year) cycle, approximately 1.1 million establishments are sampled. In the May 2022 estimates, data collected for the May 2022 panel were combined with data collected for the November 2021, May 2021, November 2020, May 2020, and November 2019 panels.

For a given panel, most sampled establishments initially receive a letter or email with instructions for reporting their data electronically. At approximately 4-week intervals, nonrespondents receive up to three additional mailings of a survey questionnaire or letter with instructions for reporting electronically. Nonrespondents may also be contacted by phone or email.

Censuses of federal and state government are collected annually:

- A census of the executive branch of the federal government and the U.S. Postal Service (USPS) is collected annually in June from the U.S. Office of Personnel Management (OPM), the Tennessee Valley Authority, and the U.S. Postal Service. Data from only the most recent year are retained for use in OEWS estimates.
- In each area, a census of state government establishments, except for schools and hospitals, is collected annually every November. Data from only the most recent year are retained for use in the estimates.

A probability sample is taken of local government establishments, private sector establishments, and state government schools.

Sampling procedures

Frame construction

The sampling frame, or universe, is a list of all in-scope nonfarm establishments that file unemployment insurance (UI) reports to the state workforce agencies. Employers are required by law to file these reports to the state where each establishment is located. Every quarter, BLS creates a national sampling frame by combining the administrative lists of unemployment insurance reports from all the states into a single database called the Quarterly Census of Employment and Wages (QCEW). Every six months, OEWS extracts the administrative data for establishments that are in scope for the OEWS survey from the most current QCEW. QCEW files were supplemented with frame files covering rail transportation (NAICS 4821) and Guam for establishments not covered by the UI program.

Construction of the sampling frame includes a process in which establishments that are linked together into multiunit companies are assigned to either the May or November sample. This prevents BLS from contacting multiunit companies more than once per year for this survey. Furthermore, the frame is matched to the 5 prior sample panels, and units that have been selected in the 5 prior panels are marked as ineligible for sampling for the current panel.

Stratification

Establishments on the frame are stratified by geographic area and industry group:

- The 2022 estimates include over 580 specified areas made up of Metropolitan Statistical Areas (MSAs) and nonmetropolitan or Balance-of-State (BOS) areas. MSAs are defined and mandated by the U.S. Office of Management and Budget (OMB). The May 2022 estimates use the August 2017 MSA definitions delineated in OMB Bulletin 17-01. Each officially defined metropolitan area within a state is specified as a substate area. Cross-state MSAs have a separate portion for each state contributing to that MSA. In addition, states may have up to six residual nonmetropolitan areas that together cover the remaining non-MSA portion of their state.
- Industry—The May 2022 estimates have about 300 industry groups defined at the NAICS 3-, 4-, 5-, or 6-digit level.

- Ownership—Schools are also stratified by state government, local government, or private ownership. Also, local government casinos and gambling establishments are sampled separately from the rest of local government.
- Size—Establishments are divided into certainty and noncertainty size classes.

Sample is taken from nonempty frame cells[1]. Frame cells are defined by State/MSA-BOS/NAICS 3-, 4-, 5-, 6-digit/ownership strata. When comparing nonempty strata between frames, there may be substantial frame-to-frame differences. The differences are due primarily to normal establishment birth and death processes and normal establishment growth and shrinkage. Other differences are due to establishment NAICS reclassification and changes in geographic location.

A small number of establishments indicate the state in which their employees are located, but do not indicate the specific county in which they are located. These establishments are also sampled and used in the calculation of the statewide and national estimates. They are not included in the estimates of any substate area. Therefore, the sum of the employment in the MSAs and nonmetropolitan areas within a state may be less than the statewide employment.

Allocation of the sample to strata

Each time a sample is selected, a 6-panel allocation of the 1.1 million sample units among these strata is performed. The largest establishments are removed from the allocation because they will be selected with certainty once during the 6-panel cycle. For the remaining noncertainty strata, a set of minimum sample size requirements based on the number of establishments in each cell is used to ensure coverage for industries and MSAs. For each nonempty frame cell, a sample allocation is calculated using a power Neyman allocation. The actual 6-panel sample allocation is the larger of the minimum sample allocation and the power allocation. To determine the current single panel allocation, the 6-panel allocation is divided by 6 and the resulting quotient is randomly rounded.

[1] About 146,000 nonempty cells in the May 2022 frame.

Two factors influence the power Neyman allocation. The first factor is the square root of the employment size of each stratum. With a Neyman allocation, strata with higher levels of employment generally are allocated more sample than strata with lower levels of employment. Using the square root within the Neyman allocation softens this effect. The second factor is a measure of the occupational variability of the industry based on prior OEWS survey data. The occupational variability of an industry is measured by computing the coefficient of variation (CV) for each occupation within the 90th percentile of occupational employment in a given industry, averaging those CVs, and then calculating the standard error from that average CV. Using this measure, industries that tend to have greater occupational variability will get more sample than industries that are more occupationally homogeneous.

Sample selection

Sample selection within strata is approximately proportional to size. To provide the most occupational coverage, establishments with higher employment are more likely to be selected than those with lower employment; some of the largest establishments are selected with certainty. The unweighted employment of sampled establishments made up 57 percent of total employment in 2022.

Permanent random numbers (PRNs) are used in the sample selection process. To minimize sample overlap between the OEWS survey and other large surveys conducted by the U.S. Bureau of Labor Statistics, each establishment is assigned a PRN. For each stratum, a specific PRN value is designated as the "starting" point to select a sample. From this "starting" point, we sequentially select the first '*n'* eligible establishments in the frame into the sample, where '*n*' denotes the number of establishments to be sampled.

Single panel weights (sampling weights)

Sampling weights are computed so that each panel will roughly represent the entire universe of establishments. Under the MB3 estimation method, sample weights are used to fit wage distribution models and wage adjustment models, but do not otherwise play a role in estimation because data are predicted for every establishment in the universe.

Federal government, USPS, and state government units are assigned a panel weight of 1. Other sampled establishments are assigned a design-based panel weight, which reflects the inverse of the probability of selection.

National sample counts

The combined sample for each set of estimates is the equivalent of six panels. The sample allocations, excluding federal government and U.S. Postal Service (USPS), for the panels in the 2022 estimates cycle are:

186,911 establishments for May 2022

187,215 establishments for November 2021

187,410 establishments for May 2021

179,303 establishments for November 2020

179,824 establishments for May 2020

179,391 establishments for November 2019

Each set of estimates includes a census of about 8,000 federal and USPS units. The combined sample size for each set of estimates is approximately 1.1 million establishments, which includes only the most recent data for federal and state government. Federal and state government units from older panels are deleted to avoid double counting.

Sample response rates

Table 2, below, contains information on unit and employment response rates and the number of viable and responding units for the 2022 estimates. The initial combined sample had approximately 1.1 million establishments.

Table 2: Response Rates for 2022 Sample

Year	Viable Units	Responding Units	Unit Response Rate	Weighted Employment Response Rate
2022	1,027,465	671,552	65.4%	62.5%

MB3 estimation methodology

Modeled occupational employment and wage estimates are made directly from a predicted population. The OEWS population is defined as the set of in-scope establishments present in the universe during the survey reference period. Each population establishment is defined by information including industry, size, ownership, and location; these are known to be strong predictors of occupational employment and wages. OEWS survey response data provide occupational employment distributions, known as staffing patterns, and wage information for a portion of the population. Models based on OEWS response data from the current panel and five previous panels predict staffing patterns and wages for the majority of the population.

Matching population units to respondent data

The prediction framework splits the population into the categories of "observed units" and "unobserved units." Observed units are defined as stable establishments with response data from the previous three years. Unit stability is defined by comparing QCEW and reported values for a number of variables, as described in the following section. For any given observed unit, occupational employment and wage data come from that unit's survey data.

Unobserved units may be nonsampled units, nonresponding units, or responding units that do not meet stability criteria. For any given unobserved unit, occupational employment and wages are predicted using data from similar responding establishments. Responding units that do not satisfy stability criteria are still usable for prediction purposes.

Exact matching—observed units

Response data from observed units are used directly in the estimates, provided a given unit is stable. The stability criteria for observed units require that reported values exactly match QCEW for the May reference panel values for 6-digit NAICS, ownership, and MSA, and be similar to population employment. Population employment for a given unit is taken as the average of the most recent May and November QCEW employment for a given reference period. The criteria for stable employment, for population employment $\boldsymbol{E_P}$ and respondent employment $\boldsymbol{E_R}$, fulfill either condition:

$$|\boldsymbol{E_R} - \boldsymbol{E_P}|/\boldsymbol{E_P} < \mathbf{0.5} \quad \text{or} \quad |\boldsymbol{E_R} - \boldsymbol{E_P}| < \mathbf{5}.$$

During data collection, states can correct the QCEW NAICS, MSA, and previous employment values if they are incorrect. These corrections are applied to the population file prior to the stability calculations so that incorrect population data do not automatically prevent a unit from being classified as stable.

Observed unit wage data from previous years are adjusted using the wage adjustment model to reflect current wage levels, and employment data are scaled to population employment levels. Partial respondent data contain information on occupational employment, but not wages. Missing wage data for partial respondents are imputed through hot deck imputation, and the units are then used as respondents. For each occupation in the partial respondent's staffing pattern that does not have complete wage data, a wage distribution is imputed from a pool of similar respondents reporting that occupation. The donor search is limited to the most recent survey panel and is initially defined by MSA/BOS, NAICS 4, size class, and, for some industries, ownership. If there are not enough donors to provide wages distributions for the partial respondents that need them, then the search criteria are loosened and the search repeated. Once a sufficiently large donor pool is found, the donor pool's wage distribution is used to prorate the recipient's reported employment in the occupation across the 12 wage intervals outlined in the "Concepts" section of this survey methods statement.

For the wage modeling process, complete nonrespondents that are missing both employment and wage data are also imputed using hot deck imputation. A single donor is used to impute the entire employment staffing pattern, followed by hot deck wage imputation with multiple donors. Once the wage modeling process is complete, the hot deck imputed data for complete nonrespondents are discarded; these units subsequently will be treated as unobserved units for estimation and will receive modeled data using the MB3 prediction process.

Some updates were made to the MB3 wage processing for the May 2022 estimates. Private sector and local government units in the May 2020 and later panels are now represented by their reported wage rates, as long as wage rate data are available for all of the establishment's employees within an occupation. If full wage rate data are not available for a given occupation, wages are sampled from a modeled wage distribution and assigned to all of the establishment's jobs in that occupation, as described in the "Wage distribution modeling and wage rate prediction" section below. This wage sampling process is also used for all private sector and local government establishments in the November 2019 survey panel, for which only interval wage data are available.

Prediction—unobserved units

The staffing patterns and wages of unobserved units in the population are predicted using data from nearest neighbor respondents. Responding units that do not pass stability criteria are no longer representative of the population cell for which they were sampled, but may be used to predict units in the cell to which they currently belong. A combination of unit matching and model-based adjustments produces these predictions for each unobserved unit. A pool of 10 nearest neighbor responding units is typically used to predict each unobserved unit. Matching is deterministic, so the predicted staffing pattern and wages of any unit of a given size, location, ownership group, and industry will be identical.

Prediction of unobserved unit staffing patterns and wages derives from a weighted sum of response data from similar units. Data from responding units that closely match the unobserved units are given more match weight relative to data from less similar units. For a given unobserved unit, a set of (typically) 10 responding units with the highest relative weights are used for the prediction.[2] Industry and establishment size as measured by total employment are the strongest predictors of staffing patterns; matches that are close in those dimensions are preferred, and a difference in either of these dimensions results in a large penalty to matching weight. Time, location, and ownership are also important predictors that play a part in finding matches; differences in any of these dimensions result in relatively smaller weight penalties

[2] There is not technically an upper bound on the number of donors. If several donors have the same weight, which puts the total number greater than or equal to 10 donors, they will all be used. For example, if there are 15 donors with the 10th highest weight, all will be used, resulting in a total of 24 donors. For the sake of explanation, we will be assuming 10 donors, which is typically the case.

than are given for industry and size differences. Matching unit weights are determined as the product of scoring functions based on each of these factors.

The match scores can be thought of as a set of penalties for deviation from a perfect match, which would have a score of 1. Matches are penalized (given lower weight) using these scores, each with a value between 0 and 1. The scoring function for each predictive factor aims to assign a score value based on the relative importance of each factor. The specific score values used in the MB3 system were evaluated using simulation studies. Various proposed scoring functions were tested to generate estimates and the best performing of these were used.

Where establishment *a* is an unobserved unit and establishment *b* is a potential match, each component of the score function accounts for differences between *a* and *b*. The score of the match is:

$$\mathrm{S}(a,b) = \mathrm{S_E}(a,b) \cdot \mathrm{S_T}(a,b) \cdot \mathrm{S_I}(a,b) \cdot \mathrm{S_O}(a,b) \cdot \mathrm{S_A}(a,b)\ where:$$

$\mathrm{S}(a,b) \leq 1$

$\mathrm{S_E}(a,b)$ – Score for difference in total employment between *a* and *b*

$\mathrm{S_T}(a,b)$ – Score for difference in time between *b* and the most recent panel

$\mathrm{S_I}(a,b)$ – Score for difference in six-digit industry between *a* and *b*

$\mathrm{S_O}(a,b)$ – Score for difference in ownership between *a* and *b*

$\mathrm{S_A}(a,b)$ – Score for difference in detailed area between *a* and *b*

A potential donor that differs in size from the unit to be predicted will be penalized for this difference. The employment component is $\mathrm{S_E}(a,b) = \left(1 - \frac{|E_a - E_b|}{E_a + E_b}\right)$, where E_a and E_b are the employment totals for the respective units. For a potential donor with 20 employees and a unit to be predicted with 15 employees, this works out to $\mathrm{S_E}(a,b) = \left(1 - \frac{|15-20|}{15+20}\right) = 0.857$.

Recently collected data are favored over data collected in previous panels. The time score component that reflects this is $\mathrm{S_T}(a,b) = 1 - \frac{p_b}{6}$, where p_b is the number of panels between the collection of data for *b*

and the reference period. A unit b observed in the current panel would have $S_T(a,b) = 1$ and a unit sampled 5 panels previously would have $S_T(a,b) = 1 - \frac{5}{6} = \frac{1}{6}$.

Donors would ideally be in the same industry or ownership group as the unit to be predicted, but may be in a similar industry or different ownership.

The score component reflecting differences in industry is $S_I(a,b) = \begin{cases} 1 & if\ industry\ matches \\ 0.25 & if\ industry\ mismatches \end{cases}$

While any difference in industry is given the same penalty, donors are chosen according to a hierarchy and therefore more similar industry matches will be used before more different matches.

The score component reflecting differences in ownership is $S_O(a,b) = \begin{cases} 1 & if\ ownership\ matches \\ 0.5 & if\ ownership\ mismatches \end{cases}$

For example, if an unobserved unit is a private school, a private school donor would have an ownership score of 1, while a public school would have an ownership score of 0.5.

Donors in the same geographical area are preferred and are treated at 4 different matching levels. Units in the same state and MSA or nonmetropolitan area have an area match score of 1, whereas units in the same state and same urban/rural status but differing MSAs receive an area match score of 0.75. Units from the same state but with differing MSAs and urban/rural status receive a score of 0.5. Units from different states receive a score of 0.25. The score component reflecting differences in area is:

$$S_A(a,b) = \begin{cases} 1 & if\ same\ MSA\ and\ state \\ 0.75 & same\ state\ and\ area\ status \\ 0.5 & if\ same\ state \\ 0.25 & otherwise \end{cases}$$

For a potential donor with 20 employees that would predict a unit of 15 employees ($S_E(a,b) = 0.857$) from 2 panels previous to the reference period ($S_T(a,b) = 0.667$), with a mismatching industry

($S_I(a,b) = 0.25$), where both are privately owned ($S_O(a,b) = 1$), and a matching state but differing MSA/nonmetropolitan area and urban/rural status ($S_A(a,b) = 0.5$), the match score is:

$$S(a,b) = 0.857 \cdot 0.667 \cdot 0.25 \cdot 1 \cdot 0.5 = 0.0714$$

Depending on the match scores of other potential donors, the unit may or may not be used in prediction.

The relative matching weight W_{b_i} of any match b_i among 10 matches, $b_1, b_2, \ldots, b_{10}$, is:

$$W_{b_i} = \frac{S(a,b_i)}{\sum_{j=1}^{10} S(a,b_j)}$$

Current panel data from the same MSA, detailed industry, ownership, and employment level as an unobserved unit will receive a weight of 1, the maximum possible weight. Potential matches are found by a hierarchical nearest neighbor search detailed in Table 3 below. All establishments with the same employment, NAICS, ownership category, state, and MSA will be predicted using the same set of donors.

An employment criterion is defined for each level such that the donor's employment must be within a certain percentage of the unobserved unit's. For example, in the first hierarchical level, the donor must be in the same state, NAICS, ownership category, and MSA of the unobserved unit to be predicted while having employment within ± 10 percent of the unobserved unit's employment. Broader industry cells use the industry groups of published OEWS estimates, which are defined at the 4-digit NAICS level in most industries, but at the 3-, 5-, or 6-digit NAICS level for a minority of industries.

If fewer than 10 potential donors are found at the first level of the hierarchy, the search proceeds through subsequent levels of the hierarchy, stopping when at least 10 suitable donors are found. If fewer than 10 donor units are available at hierarchy level 10, prediction will still proceed if at least 5 donor units are found when the search reaches the highest level. The matches with the highest scores are used for prediction.

Table 3: Hierarchical Levels of Donor Matches

Hierarchy Level	Characteristics that must match between the Prediction Cell and Responder (Donor)	Employment Criterion
1	State – NAICS – Ownership – MSA	10%
2	State – NAICS – Ownership – MSA	20%
3	State – NAICS – Ownership	10%
4	State –NAICS group– Ownership	20%
5	State – NAICS – Ownership	None
6	State – NAICS group	None
7	NAICS group– Ownership	10%
8	NAICS group– Ownership	20%
9	NAICS	None
10	NAICS group	None

Table note: "NAICS" is 6-digit NAICS. "NAICS group" is the most detailed NAICS level for which OEWS publishes estimates, generally the 4-digit NAICS level.

Employment, staffing pattern, and wage data for the 10 closest matches are used to predict the staffing pattern and wages of each unobserved unit. If the closest matches include several donors with the same weight, they will all be used, which may result in more than 10 donors. If a match differs from the unobserved unit's employment, occupational employment numbers will be scaled to the appropriate level. Occupational wage values are scaled using a wage adjustment factor if a match differs from the unobserved unit in industry, ownership, area, employment, or data collection panel. Methods for wage adjustment are discussed in the following section. For a given unobserved unit *U* and set of matches $b_i \; in \; (b_1, b_2, \ldots, b_{10})$, the predicted employment *E* for occupation *O* in wage interval *M* will be:

$$E_{UOM} = \sum_{i=1}^{10} W_{b_i} \cdot E_U \cdot \frac{E_{b_iOM}}{E_{b_i}}$$

where E_{b_iOM} is the employment in wage interval M for the occupation O in establishment b_i, E_{b_i} is the employment for establishment b_i, E_U is the employment of the unit to be predicted, and W_{b_i} is the relative matching weight of match b_i.

For establishment b_i, w_{b_iOM} represents the wage for occupation O in interval *M.* For responding establishments, w_{b_iOM} is the wage value that will represent them as observed units and as donors to

unobserved units that receive modeled data. In the May 2021 and prior estimates, responding establishments were assigned w_{b_iOM} equal to the interval *M* mean from a wage distribution modeled from the OEWS survey data. Starting with the May 2022 estimates, the establishment's reported wage rates are used as w_{b_iOM} whenever wage rate data are available for all of the establishment's employment in the occupation. If only interval wage data are available for a given establishment and occupation, then employees are assigned values of w_{b_iOM} sampled from a wage distribution, as discussed in the following section.

For the May 2021 estimates, wages for all matched donors having employment within a given occupation and interval were used to model wages for predicted employment within that occupation and interval. In many cases, fewer than 5 donors may have employment within an occupation and interval. Starting with the May 2022 estimates, a random subset of donor wages is used to predict wages for each wage interval within each occupation for unobserved units. If employment is reported for an occupation and wage interval for at least 5 donors, then it is expected that 5 donor wages will be used, but at minimum one donor wage will be used. If fewer than 5 donors are available for an occupation and wage interval, it is likely that all donor wages will be used. For each wage interval, donor wages are sampled using Poisson sampling, with a target of 5 wages for each wage interval. A systematic sample of a single unit is also taken for use in the case where no wage is sampled using Poisson sampling. The probability of selection for a given donor wage within a given wage interval in the Poisson sample is:

$$p_{UOM} = 5\frac{W_{b_i} \cdot E_{b_iOM}/E_{b_i}}{\sum(W_{b_i} \cdot E_{b_iOM}/E_{b_i})}$$

The probability of selection for systematic sampling is one fifth of the probability for Poisson sampling.

The occupational wage w predicted for unit U is derived from a weighted composite of the occupational employment of the donor units. Assuming 10 donors, this is given by:

$$w_{UOM} = \frac{\sum_{i=1}^{10} W_{b_i} \cdot \mathrm{A_O}(U, b_i) \cdot w_{b_iOM} \cdot \mathrm{I}\left(E_{b_iOM} \neq 0\right) \cdot \mathrm{I}(i \in S_{OM})}{\sum_{i=1}^{10} W_{b_i} \cdot \mathrm{I}\left(E_{b_iOM} \neq 0\right) \cdot \mathrm{I}(i \in S_{OM})}$$

Here, $\mathrm{A_O}(U, b_i)$ is the wage adjustment factor discussed below under *Model-based adjustments*. The function $\mathrm{I}\left(E_{b_iOM} \neq 0\right)$ equals 1 when the establishment's occupational employment is nonzero for a wage

interval and equals 0 otherwise. The function $I(i \in S_{OM})$ equals 1 when the wage of establishment *i* has been sampled and equals 0 otherwise.

To illustrate the prediction of an unobserved unit, suppose the unobserved unit U is a jewelry store in a medium-sized town. The staffing pattern is predicted for each wage interval of each occupation in the nearest donors. Detailed below is the prediction for retail salespersons in wage interval C. Suppose that of the 10 nearest respondents available, eight had employment of retail salespersons in wage interval C. Of those eight, four are other jewelry stores of similar size in the same MSA and most recent survey panel. Three are other jewelry stores in the same MSA with larger differences in size. Of those three, one is from one panel back. The last unit is in the same state and most recent panel, of a similar size, but in a different industry and MSA. To predict the employment and wage for retail salespersons in unobserved unit U, wage interval C, we use the example data and calculations in table 4.

Table 4: Example Data for Predicting Employment

Matching Unit: i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Unit *i* relative match weight: W_{b_i}	0.116	0.116	0.116	0.113	0.113	0.110	0.107	0.097	0.088	0.022
Unit *i* employment in interval *C*, occupation *O*: $E_{b_i OC}$	4	3	0	2	3	7	8	0	3	11
Unit *i* total employment: E_{b_i}	21	21	21	22	20	19	18	21	25	21
Unobserved unit total employment: E_U	21	21	21	21	21	21	21	21	21	21
Unit *i* occupational employment ratio in interval *C*, occupation *O*: $\frac{E_{b_i OC}}{E_{b_i}}$	0.190	0.143	0	0.091	0.150	0.368	0.444	0	0.120	0.524
Wage adjustment factor: $\mathbf{A_O}(U, b_i)$	1	1	1	1	1	0.9	1	1.03	1.03	1.1
Unit *i* wage in interval *C*, occupation *O*: $w_{b_i OC}$	13.25	13.25	N.A.	13.25	13.25	13.25	13.25	N.A	13.05	13.25
Computed unit *i* Employment Share: $W_{b_i} \cdot E_U \cdot \frac{E_{b_i OC}}{E_{b_i}}$	0.463	0.348	0	0.216	0.356	0.850	0.998	0	0.222	0.242
Poisson Sampling Indicator: $\mathrm{I}(i \in S_{OM})$	0	1	0	0	1	0	1	0	1	0
Computed unit *i* Wage Share: $\frac{W_{b_i} \cdot \mathbf{A_O}(U, b_i) \cdot w_{b_i OC} \cdot \mathbf{I}(E_{b_i OC} \neq \mathbf{0}) \cdot \mathrm{I}(i \in S_{OM})}{\sum_{i=1}^{10} W_{b_i} \cdot \mathbf{I}(E_{b_i OC} \neq \mathbf{0}) \cdot \mathrm{I}(i \in S_{OM})}$	0	3.625	0	0	3.531	0	3.344	0	2.790	0

Note: data do not correspond to existing establishments or weights.

Summing the third-to-last line ("Computed unit *i* Employment Share") of Table 4 yields predicted employment of retail salespersons O in wage interval C for establishment U:

$$E_{UOC} = \sum_{i=1}^{10} W_{b_i} \cdot \frac{E_U}{E_{b_i}} \cdot E_{b_i OC}$$

$$= 0.463 + 0.348 + 0 + 0.216 + 0.356 + 0.850 + 0.998 + 0 + 0.222 + 0.242$$

$$= 3.695$$

Although this does not add up to a whole number, for estimation purposes it is reasonable. Summing the last line of table 4 yields the predicted wage of retail salespersons *O* in wage interval *C* for establishment *U*:

$$w_{UOC} = \sum_{i=1}^{10} \frac{W_i \cdot \mathrm{A_O}(U, b_i) \cdot w_{b_i OC} \cdot \mathrm{I}\left(E_{b_i OC} \neq 0\right) \cdot \mathrm{I}(i \in S_{OM})}{\sum_{i=1}^{10} W_i \cdot \mathrm{I}\left(E_{b_i OM} \neq 0\right) \cdot \mathrm{I}(i \in S_{OM})}$$

$$= 0 + 3.625 + 0 + 0 + 3.531 + 0 + 3.344 + 0 + 2.790 + 0$$

$$= \$13.29$$

When this process is completed for all occupations and wage interval levels observed in the match units, the predicted wage and employment profile of establishment *U* can be used for estimation.

Wage parameters

For units with interval wage data, observed units from earlier survey panels, and unobserved units to be predicted, wage data require additional adjustments before being used to calculate wage estimates. This wage data processing uses both wage rates and the wage interval groups shown in Table 1 above.

Using interval data to compute mean wage estimates requires that a wage value be assigned to each employee. MB3 wage estimates use sampled wage rates that are computed using log-normal models fit to each panel of OEWS wage data, aggregated by occupation group and area group.

Predicting unobserved units also requires adjusting wages in the donor units to current local dollars for the unobserved units. For example, let's suppose an interior design firm (NAICS 541410) in a large metropolitan area and surveyed in a previous survey panel contributes to the wage prediction for an industrial design firm (NAICS 541420) in a small metropolitan area. Occupational wages will differ between these firms due to geography, industry, and time effects. Thus, wages from the first unit must be adjusted with these factors in mind to give a reasonable prediction of the second unit. A fixed effect linear

regression model, fit to observed unit data, is the basis for these adjustments. Wages for observed units collected in earlier survey panels are also updated to the reference date using a regression model.

The wage distribution model and wage adjustment model both use weighted least squares regression to estimate model parameters. Benchmarked sample weights are used in this process, such that weighted employment totals for the current panel will equal QCEW frame values for each industry, state, MSA, and size subgroup.

In MB3, benchmarking factors are used only to adjust data for the purposes of model fitting and are not used directly for estimation. Benchmarking for OEWS uses the average of the May 2022 and November 2021 QCEW employment to adjust the weighted reported occupational employment and improve the accuracy of the sampled wage rates and wage adjustment models. The ratio estimation process is carried out through a series of four hierarchical employment ratio adjustments. The ratio adjustments are also known as benchmark factors (BMFs). The BMFs are calculated for the cells defined at each of the following hierarchy levels:

Level	Area	Industry	Size	Ownership
1	MSA/BOS	NAICS 3/4/5/6 digits	1-19, 20-49, 50-249, 250+	
2	State	NAICS 3/4/5/6 digits		
3	State	NAICS 3 digits		For hospitals, schools, gambling establishments, and casino hotels
4	State	NAICS 2 digits		

For each establishment, a BMF is generally calculated by finding the ratio of QCEW employment (average of May 2022 and November 2021) to weighted cell employment for the hierarchy level. There is a universal maximum and minimum BMF value to which the BMF will be set if it is higher than the maximum or lower than the minimum. The second, third, and fourth BMF hierarchy levels are computed to account for inadequate coverage of the universe employment—for example, if an establishment is in a first-level hierarchy cell with no other establishments. The BMFs are dependent upon the establishment's previous hierarchy levels BMFs. A final benchmark factor is calculated for each establishment as the

product of its four hierarchical benchmark factors. A benchmark weight value is then calculated as the product of the establishment's six-panel combined sample weight and final benchmark factor.

Wage distribution modeling and wage rate prediction

Wage rate values assigned to interval count data are produced using modeled wage distributions. Wage distributions are modeled for each panel using only weighted data from that panel to represent the population. Occupation and locality are the strongest predictors of wages and may cause substantial differences in wage levels between establishments. To provide greater homogeneity within the data, occupations and areas with similar median wages are aggregated into groups.

We assign occupation group codes to all occupations with median wages in a given wage interval, and likewise assign group codes to all geographic areas with similar median wages. Then, all data with a given ownership status, occupation group, and area group are pooled together for modeling a wage distribution function. The units within an occupation-area-ownership group are not necessarily related in any way other than the wage interval that the median falls into.

Wage distribution modeling incorporates reported wage rate data (specific wages of each employee) from private and local government establishments. The wage distributions for each group are modeled by a log normal model fit using a log-likelihood expression that incorporates both wage rate and wage interval data. For previous estimates, wage interval means derived from the wage distribution function represented wage rates for all employees that were processed as intervalized counts. Starting with the 2022 estimates, wage rates were sampled from the appropriate modeled distribution for employees reported as wage interval counts.

The assignment of occupation wage groups uses single panel sample weights and reported employment levels within wage intervals to compute the national wage distribution for each detailed occupation, and then determine into which interval the median wage for that occupation falls. This determines the wage occupation group for every six-digit occupation. To be specific, we calculate occupation-specific employment in each of the twelve wage intervals in panel p:

$$\hat{E}_{ob_p p} = \sum_{e \in R_p} w_{ep} \times E_{ob_p ep}$$

where R_p represents the set of panel p OEWS sampled units, w_{ep} is the sample weight for establishment e in panel p, and E_{ob_pep} is the reported level of employment in occupation o at establishment e in wage interval b_p and panel p.[3] We then calculate total occupation-specific employment:

$$\hat{E}_{op} = \sum_{b_p} \hat{E}_{ob_pp}$$

and compute the relative employment shares by wage interval:

$$\hat{s}_{b_p|o,p} = \hat{E}_{ob_pp} \Big/ \hat{E}_{op}$$

We then compute cumulative employment shares:

$$\pi_{b_p|o,p} = \sum_{b \le b_p} \hat{s}_{b_p|o,p}$$

The detailed occupation is mapped into the aggregate occupation O in the lowest wage interval that contains at least 50 percent of the detailed occupation's cumulative employment:

$$\pi_{O-1|o,p} < 0.5 \le \pi_{o|o,p}$$

so that each aggregate occupation corresponds to a wage interval.

Typically, there are either 11 or 12 aggregate occupations corresponding to the various wage intervals.[4] For example, tax preparers, substitute teachers, and fast food cooks all have median wages in interval C, so they are grouped together in occupation group C, while architectural and civil drafters, actors, and

[3] The wage interval is indexed by p.

[4] There are 12 wage intervals, but each wage interval is not necessarily assigned an aggregate occupation.

construction and building inspectors all have median wages in interval F and are therefore grouped together into occupation group F.

Similarly, we compute the wage distribution for each detailed geographic area (across all occupations) and then determine in which interval the median wage for that area would fall. This determines the aggregate area for every detailed MSA or BOS area. To be specific, we calculate area-specific employment in each of the twelve wage intervals b_p in the current panel:

$$\hat{E}_{ab_pp} = \sum_o \sum_{e \in R_{ap}} w_{ep} \times E_{ob_pep}$$

where R_{ap} represents the set of panel p OEWS sampled units in area a, w_{ep} represents the sampling weight for establishment e in panel p, and E_{ob_pep} is the reported level of employment in occupation o at establishment e in wage interval b_p and panel p.[5] We then calculate total area-specific employment:

$$\hat{E}_{ap} = \sum_{b_p} \hat{E}_{ab_pp}$$

and compute the relative employment shares by wage interval:

$$\hat{s}_{b_p|a,p} = \hat{E}_{ab_pp} \Big/ \hat{E}_{ap}$$

We then compute cumulative employment shares:

$$\pi_{b_p|a,p} = \sum_{b \le b_p} \hat{s}_{b_p|a,p}$$

[5] The wage interval is indexed by p.

The detailed area is mapped into the aggregate area A in the lowest wage interval that contains at least 50 percent of the detailed area's cumulative employment:

$$\pi_{A-1|a,p} < 0.5 \leq \pi_{A|a,p}$$

so that each aggregate area corresponds to a wage interval.

Typically, there are only three or four aggregate areas, corresponding to interval C, D, E, or F. For example, the median wages in San Francisco, CA, and Boston, MA, fall into wage interval F and these areas are therefore grouped together in area group F, while the median wages in Chicago, IL, and Atlanta, GA, fall into wage interval E and these areas are therefore grouped together in area group E.

Now that we have calculated the separate aggregate occupations and aggregate areas, we combine them to create aggregated area and occupation groups within a wage interval. These aggregate occupation-areas are necessary to correctly adjust the parameters of the log-normal model and subsequently predict local sampled wage rates.

For every possible aggregate occupation-area, denoted as OA, we compute the single panel sample-weighted employment levels for each wage interval:

$$\hat{E}_{OAb_p p} = \sum_{o \in O} \sum_{e \in R_{Ap}} w_{ep} \times E_{oeb_p p}$$

where R_{Ap} is the panel p sample in aggregate area A, w_{ep} represents the sampling weight for establishment e in panel p, and $E_{ob_p ep}$ is the reported level of employment in occupation o at establishment e in wage interval b_p and panel p.[6] In general, there will be a limited number of aggregate occupation-area groups, typically between 33 and 48.

[6] The wage interval is indexed by p.

For example: nurses and paralegals are in occupation group D for a given panel, while doctors and lawyers happen to be in occupation group G. Their employers, a hospital and a law firm, are in different metropolitan areas, but both areas are in area group C. Both employers are also privately owned with ownership code 5. The data for nurses from the hospital and paralegals from the law office are pooled with other data from the same occupation-area-ownership combination to estimate wage group DC5, while the data for doctors in the hospital and lawyers in the law firm are pooled with other data to estimate wage group GC5.

A log-normal model is fit to these aggregated-occupation-by-aggregated-area cells. A maximum likelihood estimator and the sample-weighted employment sums from the current sample are used to estimate the two parameters of the lognormal model for wage *w*, occupation *O*, and area *A,* which falls into a wage interval:

$$\ln(w_{AO}) \sim \mathrm{N}(\mu_{AO}, \sigma^2_{AO})$$

Local sampled wage rates are predicted using these wage distribution parameter estimates. All data from occupation-area-ownership group OA are used to fit a log-normal model. We then sample a wage for the appropriate wage interval from the wage distribution model. For example, say a paralegal's reported wage falls into wage interval E, while their occupation and geographic location fall into occupation-area group DC. This paralegal will be assigned a wage sampled from the log-normal model within the specific occupation-area aggregate group. Each paralegal in this area reported in interval E will be independently assigned an interval E wage rate that was sampled from the distribution modeled for occupation-area group DC.

This process converts all wage interval data to wage rate values. These sampled wage rates, along with usable reported wage rates, directly define wages for all respondents. Respondent wages are adjusted, if needed, to define wages for unobserved units. For each interval containing a state or federal minimum wage, the sampled wage rate is taken from the range above the minimum wages.

Wage regression modeling

Using observed units to predict unobserved units relies on similarity between the units. Wage adjustment is necessary if the unobserved unit differs from donor units in industry, size, location, or time of data

collection. Sample response data from the current and previous two years are used to fit fixed effect linear regression models for wage corrections. Coefficients are determined using maximum likelihood estimation over data from the six panels. For a given occupation O, the model is of the form:

$$\ln(w_{OT}) \sim \beta_{\mathrm{O}} + \beta_{\mathrm{IH}} + \beta_{\mathrm{A}} + \beta_{\mathrm{S}} \cdot \mathrm{E} + \beta_{\mathrm{TM}} + \epsilon$$

w – wage

β_{O} – occupation effect across about 850 detailed occupations

β_{IH} – industry-ownership combined effects across about 1,100 detailed NAICS and ownership combinations

β_{A} – area effect across about 480 detailed areas

β_{S} – size effect linear coefficient

E – total establishment employment

β_{TM} – time effect computed independently for each of 22 major occupational groups

Model-based adjustments: wage aging and cell-level adjustments

Aging factors, which provide adjustments for changes to occupational wages over time, and locality adjustments are both computed directly using wage regression model parameters. All direct match units are separately aged according to a factor based on the combination of year and SOC major group. All donors, including unstable units, are independently adjusted to account for the year and SOC major group combination, detailed occupation, industry, ownership, and size of the unit to be predicted.

Suppose unit *b* is selected as a donor for unit *a.* Unit *b* is one of typically 10 donors for unit *a* and might come from one or more panels back, in which case unit *b*'s wage data are adjusted to match local current dollars for unit *a*. The adjusted donor wage for occupation *O* in unit *a* based on adjusted unit *b* data is:

$$\widetilde{w}_{bO} = w_{bO} \cdot A_O(a, b)$$

Where:

$$A_O(a,b) = \exp\left(\frac{\beta_{\mathrm{O}}(a) + \beta_{\mathrm{IT}}(a) + \beta_{\mathrm{A}}(a) + \beta_{\mathrm{S}} \cdot \mathrm{E_a} + \beta_{\mathrm{MP}}(a)}{\beta_{\mathrm{O}}(b) + \beta_{\mathrm{IT}}(b) + \beta_{\mathrm{A}}(b) + \beta_{\mathrm{S}} \cdot E_b + \beta_{\mathrm{MP}}(b)}\right)$$

Estimates

Occupational employment and wage estimates are computed using observed data and predicted data for the population of about 8 million[7] units. Predicted data are available for each unit of the population, so employment estimates are computed by summing employment within an estimation cell and wage estimates are computed by dividing summed wages by total employment for an estimation cell.

Occupational employment estimates

Estimates of occupational employment totals are computed by summing all employment counts of a given occupation over the modeled population data. Estimates are made over area, industry, and ownership. For occupation *o*, where unit *i* is any establishment in cell *c*, the occupational employment estimate is:

$$\hat{X}_{o,c} = \sum_{i \in o,c} x_{i,o}$$

Hourly wage rate estimates

Mean hourly wage is calculated as the total hourly wages for an occupation divided by its total estimated population employment. Wage rate information is available for federal executive branch and USPS employees, most state government employees, and most private sector and local government employees. All other wage data are placed in wage interval ranges and converted to local hourly wages for each employee. These local hourly wages are predicted using adjusted estimates of local sampled wage rates, and thereafter are treated as wage rate data. Mean wage is calculated as a sum of the hourly wage for each employee in a cell divided by the total number of employees in a cell. Employees *E* in a given occupation and wage interval at a single establishment will all have the same predicted wage *w*. For establishment *i*, wage range *r*, and occupation *o* in cell *c*, the computation is as follows:

$$\widehat{w}_{c,o} = \frac{\sum_{i \in c,o} \sum_{r} x_{iro} \cdot w_{iro}}{\sum_{i \in c,o} x_{iro}}$$

Percentile wage rate estimates are computed directly from the predicted population using the empirical distribution function with averaging, which is available in many statistical packages.

---

[7] Millions of units per year: 8.32 in May 2022.

Annual wage rate estimates

These estimates are calculated by multiplying mean or percentile hourly wage rate estimates by a "year-round, full time" figure of 2,080 hours (52 weeks x 40 hours) per year for most occupations. These estimates, however, may not represent mean annual pay should the workers work more or less than 2,080 hours per year.

Although both annual and hourly wage estimates are produced for most occupations, only annual or only hourly wages are produced for some occupations. For example, some workers such as teachers, pilots, and flight attendants are paid on an annual basis but do not work the usual 2,080 hours per year. For these workers, survey respondents report annual wages. Since the survey does not collect the actual number of hours worked, hourly wage rates cannot be derived from annual wage rates with any reasonable degree of confidence. As a result, only annual wage estimates are produced for these occupations. Other workers, such as some entertainment workers, are paid hourly rates, but generally do not work 40 hours per week, year round. For these workers, only hourly wage estimates are produced.

Variance estimation

Variances for both mean wage estimates and occupational employment estimates are computed using the "bootstrap" replication technique. Many weights may be associated with a given respondent in MB3 estimates because that respondent may be used to predict multiple unobserved units. This presents problems for many approaches to computing sampling variances. However, bootstrap sample replication is amenable to this design because the full MB3 estimation system may be applied to each replicate sample. Studies that were performed using simulated data informed decisions on the specifics of the bootstrapping approach used here and the number of replicates needed for estimates to converge.

The MB3 variances are computed over 300 bootstrap sample replicates. Each set of replicate estimates is based on a subsample of the full sample and includes model fitting as well as population prediction based on this subsample. The subsample is drawn from the full sample using a stratified simple random sample with replacement design, where the size of the subsample is equal to the size of the full sample. By sampling with replacement, we are up-weighting some sampled units by including them more than once in the subsample and down-weighting others by not including them at all. MB3 selects six independent subsamples, one from each of the six semiannual OEWS panel samples. The stratification plan is the

same used for drawing the full sample, where strata are defined by state, MSAs, aggregate NAICS industry, and ownership for schools and hospitals.

Subsampling only occurs for the noncertainty sample units. All certainty units from the full sample are used in every replicate's bootstrap sample. Some strata may only contain a single noncertainty unit, for which a variance cannot be computed. These are referred to as 1-PSU strata. A collapsing algorithm combines these 1-PSU strata with other like strata to ensure that two or more noncertainty sample units are present in a particular stratum. The collapsing is by the hierarchy detailed in Table 5.

Table 5: Hierarchical Definitions for Collapsing 1-PSU Strata

Hierarchy Level	Collapse
1	Panels*: (0,1), (2,3), and (4,5)
2	Panels*: (0,1,2), and (3,4,5)
3	Panels*: (0,1,2,3,4,5)
4	MSAs
5	Allocation NAICS (A_NAICS)
6	Nationally

* Panels are labeled 0 to 5, where 0 corresponds to the May 2022 panel

The probability of selection for resampling a given unit is proportional to that unit's share of stratum employment. This ensures that bootstrap sample replicates yield unbiased estimates of the full sample's estimates. Sampling variance estimates obtained through these methods do not use the same probabilities used in selection of the full sample, which presents a possible source of error. Studies indicate that these estimates are approximately unbiased and perform well in estimating sampling variance.

The six replicate subsamples are combined for calculating MB3 replicate estimates. Single panel sample weights for the most recent panel are retained for computation of wage distribution model parameters and the wage adjustment factors in each replicate. All matching, wage parameter, and estimation methods

described previously are used with each 6-panel bootstrap subsample to create occupational employment and wage replicate estimates for every estimation cell. This process is repeated to create 300 sets of replicate estimates. For every estimation cell in which OEWS calculates an estimate, there are occupational employment and mean wage estimates based on the full OEWS sample, as well as 300 occupational employment and mean wage replicate estimates each based on a different bootstrap subsample. The variance estimates for the occupational estimates based on the full sample are calculated by finding the variability across the occupational replicate estimates. The below formula calculates the bootstrap variance estimates:

$$v_{BS}(\hat{\theta}_{j,D}) = \frac{1}{(300-1)} \sum_{b=1}^{300} \left(\hat{\theta}_{j,D}^{(b)} - \hat{\theta}_{j,D} \right)^2$$

where

$\hat{\theta}_{j,D}$ = occupational estimate (employment or mean wage) for occupation *j*, within estimation domain *D*, based on **full** sample

$\hat{\theta}_{j,D}^{(b)}$ = occupational replicate estimates (employment or mean wage) for occupation *j*, within estimation domain *D*, based on the bootstrap subsample for replicate b

Reliability of the estimates

Estimates developed from a sample will differ from the results of a census. An estimate based on a sample survey is subject to two types of error: sampling and nonsampling error. An estimate based on a census is subject only to nonsampling error.

Nonsampling error

This type of error is attributable to several causes, such as errors in the sampling frame; an inability to obtain information for all establishments in the sample; differences in respondents' interpretation of a survey question; an inability or unwillingness of the respondents to provide correct information; and errors made in recording, coding, or processing the data. Explicit measures of the effects of nonsampling error are not available.

Sampling error

When a sample, rather than an entire population, is surveyed, estimates differ from the true population values that they represent. This difference, the sampling error, occurs by chance and depends on the particular random sample used in a survey. Sampling error is characterized by the variance of the estimate or the standard error of the estimate (square root of the variance). The relative standard error is the ratio of the standard error to the estimate itself.

Estimates of sampling variability for occupational employment and mean wage rates are provided for all employment and mean wage estimates to allow data users to determine if those statistics are reliable enough for their needs. Sample estimates from a given design are said to be unbiased when an average of the estimates from all possible samples yields the true population value. Empirical studies support that MB3 methods provide accurate estimates of sampling variability.

Estimated standard errors should be taken to indicate the magnitude of sampling error only. They are not intended to measure nonsampling error, including any biases in the data. Particular care should be exercised in the interpretation of small estimates or of small differences between estimates when the sampling error is relatively large or the magnitude of the bias is unknown.

Quality control measures

Several edit and quality control procedures are used to reduce nonsampling error. For example, data submissions are checked for consistency. Follow-up emails, postal mailings, and phone calls are sent out to nonresponding establishments to improve the survey response rate.

The OEWS survey is a federal-state cooperative effort that enables states to conduct their own surveys. A major concern with a cooperative program such as OEWS is to accommodate the needs of BLS and other federal agencies, as well as state-specific publication needs, with limited resources while simultaneously standardizing survey procedures across all 50 states, the District of Columbia, and the U.S. territories. Controlling sources of nonsampling error in this decentralized environment can be difficult. One important quality control tool used by the OEWS survey is a computerized survey data management system. This was developed to provide a consistent and automated framework for survey processing and to reduce the workload for state, regional, and national office analysts.

To ensure standard sampling methods in all areas, the sample is drawn in the national office. Standardizing data processing activities, such as validating the sampling frame; allocating and selecting the sample; refining mailing addresses; designing and updating letters, emails, and questionnaires; conducting electronic review; producing management reports; and calculating estimates, have resulted in the overall standardization of the OEWS survey methodology. This has reduced the number of errors on the data files as well as the time needed to review them.

Other quality control measures used in the OEWS survey include:

- Follow-up email, postal mail, and telephone solicitations of nonrespondents, especially critical or large nonrespondents
- Review of data during collection to verify its accuracy and reasonableness
- Adjustments for atypical reporting units on the data file
- Validation of unit matching and donor profiles

Confidentiality

BLS has a strict confidentiality policy that ensures that the survey sample composition, lists of reporters, and names of respondents will be kept confidential. Additionally, the policy assures respondents that published figures will not reveal the identity of any specific respondent and will not allow the data of any specific respondent to be inferred. The most relevant statute which governs BLS confidentiality is the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). Each published estimate is screened to ensure that it meets these confidentiality requirements. To further protect the confidentiality of the data, the specific screening criteria are not listed in this publication. For additional information regarding confidentiality, please visit the BLS website at www.bls.gov/bls/confidentiality.htm.

Data presentation

Included are cross-industry data for the United States as a whole, for individual U.S. states, and for metropolitan and nonmetropolitan areas, along with national industry-specific estimates by 2-digit, 3-digit, most 4-digit, and some 5- and 6-digit NAICS levels. Available data include estimates of

employment; annual mean wages; 10th, 25th, 50th (median), 75th, and 90th percentile wages; and relative standard errors (RSEs) for the employment and mean wage estimates.

Uses

For many years, the OEWS survey has been a major source of detailed occupational employment data for the nation, states, and areas, and by industry at the national level. This survey provides information for many data users, including individuals and organizations engaged in planning vocational education programs, higher education programs, and employment and training programs. OEWS data are used to prepare information for career counseling, for job placement activities performed at state workforce agencies, and for personnel planning and market research conducted by private enterprises. OEWS data also are used by the Department of Labor's Foreign Labor Certification (FLC) program, which sets the rate at which workers on certain work visas in the United States must be paid.