UPDATING OCCUPATIONAL COMPENSATION SURVEY PROGRAM DATA USING THE EMPLOYMENT COST INDEX

Jason Tehonica, Bureau of Labor Statistics
2 Massachusetts Ave., N.E., Room 3160, Washington, DC 20212

Key Words: Complex Survey Sampling; Distribution Function; Regression

The Occupational Compensation Survey Program (OCSP), in addition to estimating occupational wages for Locality Pay, constructs national estimates from a probability selection of establishments stratified geographically and by industrial activity. The primary sampling units (PSUs) are typically Metropolitan Statistical Areas. The OCSP collects data, principally wages, on all incumbents within the selected establishments for a predefined list of occupations and publishes estimated means, quartiles, and distributions. The original design included surveying all PSUs every year, but, because of budget constraints, and a desire to minimize the collection and respondent burden, the design was altered to include surveying smaller PSUs only in alternate years. National estimates can be produced annually by updating the previous year's wage data in PSUs where no data is currently collected, using the Employment Cost Index (ECI).

The Employment Cost Index gives a year by year measure of the cost to employers of wages and salaries for all occupations and establishments in both the private non-farm sector and State and local governments, thus permitting analysis of labor cost changes for a major portion of the U.S. economy. The ECI survey is different from the OCSP in that it is designed to make estimates for major occupational groups instead of specific occupations. Nevertheless, like the OCSP, the ECI calculates estimates on a national basis. We considered it a reasonable choice of an index by which to measure the amount that wages might have changed from a preceding year to the current year.

To understand the update process, it is important to understand the survey design of the OCSP as it pertains to the national estimates. The current design entails three types of surveyed areas: certainty metropolitan areas, non-certainty metropolitan areas, and non-certainty, non-metropolitan (non-met) areas. The certainty areas are predominantly primary metropolitan statistical areas (PMSAs) where wage data is collected every year (for example, New York, NY and Los Angeles, CA). These areas consist of cities with relatively large private, non-agricultural employment. Because of the OCSP's intent to estimate wages in cities with relatively large federal government employment, most of these certainty areas coincided with areas of interest in the Locality Pay Program, which is intended to compute estimates of wages in selected areas to be used to adjust federal salaries in accordance with the Federal Employees Pay Comparability Act of 1990[1]. The non-certainty areas were selected to supplement the certainty areas in an attempt to generate national estimates. These non-certainty areas are smaller metropolitan areas (for example, Danbury, CT and Hartford, CT) and non-met counties (for example, McKean County, PA and Mason County, WV) that are paired in such a way that one PSU from each pair is surveyed only in alternating years. The non-certainty area that is not surveyed in a given year will have its previous year's wage data updated using the ECI. The non-certainty areas are paired by trying to match areas by size, geographic region, and industrial mix.

As an example of how the update process works, let's say that two of the non-certainty metropolitan areas in the OCSP that are paired are Danbury, CT and Hartford, CT. If we were computing national estimates for 1995, Hartford may be surveyed for that year and then the Danbury data from 1994 would be updated with the ECI, and this updated data would be used to compute the national estimates. Then in 1996, Danbury would be surveyed and the wage data collected in 1995 from the Hartford survey would be updated with the appropriate ECI factor and used for the 1996 national estimates. The non-met areas would be treated in the same manner.

A study was carried out on earlier wage data to achieve an understanding of the impact of updating on estimates. Since the OCSP is new, it could not supply enough data to test the updating process over a period of years. Therefore, we decided to use archived data from BLS's Area Wage Survey (AWS) Program, a predecessor to the OCSP, and mimic the OCSP design. The 70 areas that comprised the AWS program were divided into certainty areas and non-certainty areas. The non-certainty areas were divided into matched halves. In a given year, the survey areas consisted of the certainty areas, the "surveyed" non-certainty areas, and the "deselected" or non-surveyed, non-certainty

[1] Berry Newman, Constance (November 9, 1990); *Major Features of the Federal Employees Pay Comparability Act of 1990*

areas. These "deselected" areas would have their data updated from the previous year using the ECI.

The study was designed to test the update process on the national estimates for the years 1982 through 1985. These years were chosen because we wanted two years with relatively stable economies (1984 and 1985) and two years that portrayed more volatile changes (1982 and 1983). We also wanted to test a cross section of occupations to ascertain whether the updating process performed differently across major occupational groups. The study involved 35 occupations, spanning the professional, technical, clerical, and blue collar occupational groups.

One question was which component of the ECI would best approximate the change from one year to the next for the 35 occupations involved in the study. We primarily concerned ourselves with two approaches. Each approach represented the percent change in wages and salaries for private industry workers for the twelve months ending in June of the year for which the data was being updated. The first approach used the ECI factor of white-collar occupations excluding sales to update for all occupations. This index was chosen because of its prior use to update wage data for the White Collar Pay (WCP) Program.[2] The second approach used a different ECI factor for each major occupational group. Professional occupations were updated with the "White-Collar Less Sales" factor; technical with the "Professional Specialty and Technical Occupations" factor; clerical with the "Administrative Support Including Clerical Occupations" factor; and blue-collar jobs with the "Blue-Collar Occupations" factor.

To illustrate the actual updating process, we look at the national estimates for 1983 using the "White-Collar Less Sales" ECI factor. First we deleted the wage data for half of the non-certainty areas to simulate the areas that would not be surveyed in 1983. The 1982 wage data for those areas was then updated to reflect data that would have been collected in 1983. For the twelve months ending in June 1983 the "White-Collar Less Sales" ECI factor was 1.059, which indicates a 5.9% change from June 1982 to June 1983. This factor was applied to each wage record in the 1982 data for those areas that would not be surveyed in 1983. This updated data was then combined with the actual collected 1983 data from the 33 certainty areas and the remaining half of the non-certainties. The national wage estimates for 1983 were then computed from this group of data. For example, the mean wage is given by:

$$\hat{\mu} = \frac{\sum_{C} w_{hi} \sum_{j} x_{hij} + \sum_{N1} w_{hi} \sum_{j} x_{hij} + \sum_{N2} w_{hi} \sum_{j} \hat{x}_{hij}}{\sum_{C} w_{hi} E_{hi} + \sum_{N1} w_{hi} E_{hi} + \sum_{N2} w_{hi} E_{hi}}, \text{ where:}$$

w = area/establishment weight,
x = earnings for a worker in an occupation,
E = employment for each occupation,
h = area stratum,
i = establishments stratum,
j = occupational wage record,
C = certainty areas,
N1 = surveyed non-certainty areas,
N2 = non-surveyed non-certainty areas, and
$\hat{x}$ = earnings updated using ECI and previous year's data.

These estimates could then be compared to wage estimates from the actual collected 1983 data from all 70 areas (33 certainties and 37 non-certainty areas).

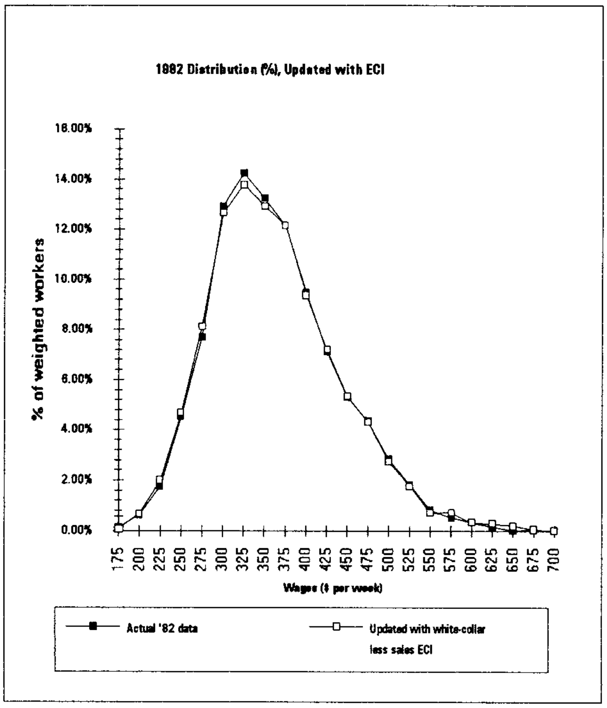
A measure of the deviation for the estimated means, as well as estimated quartiles, was calculated by using the relative difference (rd) of the updated data estimate from the full data estimate. *(rd = (updated estimate - full sample estimate) / full sample estimate)* If this difference were to be much greater than the amount of variation normally present in the full sample data estimates, then the update methodology would have to be reconsidered. For the estimated means, a relative difference of greater than 1% was considered to be worrisome since the relative error in national estimates is typically about 1%. Our results showed that a majority of occupations studied had an rd of less than 1% for their estimated means. Study of the quartiles suggested that although updating was less effective than the means, a majority of occupations had rds less than 1% as well.

Probably the most important aspect of the update process was to see how it would affect the distribution of workers over the range of wages. This is because of the Bureau's desire to publish more data on the positional statistics of the distributions and also to develop variance estimates for these positional statistics.

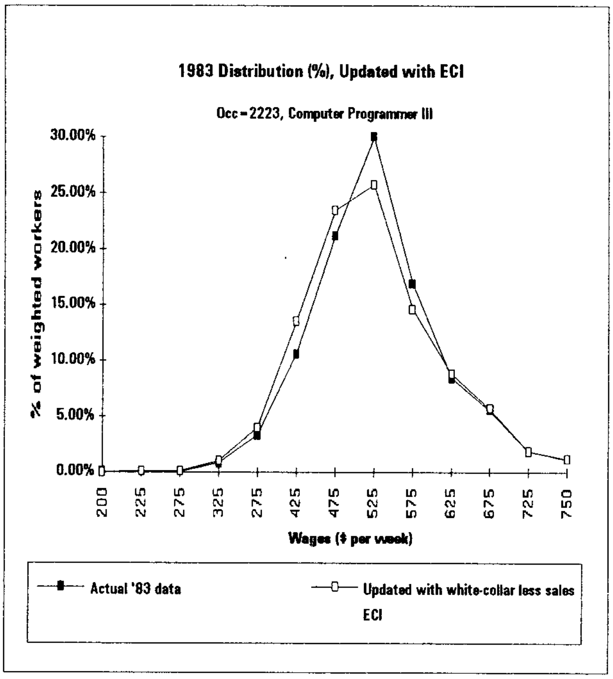
To capture the effect on the distributions, the distribution of workers across wage ranges for the actual data and the updated data was graphed for selected occupations. The actual data is the data collected from all areas in that year, that is, data from all the PSUs instead of treating half of the smaller PSUs as deselected. The updated data is the actual collected data from the certainties and half of the non-certainties combined with the previous year's wage data, updated with the appropriate ECI factor, for the other half of the non-certainties. The graph below is

[2] Burdette, Terry (May 21, 1991); *1991 WCP Estimates Revisited*

the distribution for the Secretary level IV job which we collect wage data for.



The graph of the actual, collected data and the graph of the estimates that included some data updated with the ECI are very similar. Graphing the wage data for all jobs in the study showed that updating had little effect on the shape of the distribution for most jobs. There were a few occupations where the distribution was slightly shifted when updated data was used such as the graph for the Computer Programmer Level III job shown below. But the degree to which the distributions were affected was still in a range acceptable for the purposes of the study.



To determine an appropriate measure to use for the deviations of the distributions, we wanted to find a statistic that was similar to the relative difference used for the means and quartiles. We considered several test statistics but focused on one in particular, $\max_K(\left|\sum_{k=1}^{K} p_{k,update} - \sum_{k=1}^{K} p_k\right|)$, where K is the number of wage intervals and p is the proportion of total workers in that interval. This is the maximum deviation of the distribution function and corresponds to a Kolmogorov-Smirnov[3] test statistic. It is important to note that actually performing the Kolmogorov-Smirnov test of hypothesis is inappropriate here because the two samples have data in common. Other statistics we considered corresponded to a chi-squared goodness of fit test and the Cramér-von Mises[3] test. Our results suggested that for most of the occupations in our study, the update process has no sharp effect on the distributions.

To illustrate how the graphical analysis translates to the test statistic that was used, we can compare the graphs from above with the value for the maximum deviation of the distribution function that was calculated. The table below shows a sample of occupations contained in the study and their corresponding test statistic value.

[3] Durbin, J. (1980); *Distribution Theory for Tests Based on the Sample Distribution Function*; Society for Industrial and Applied Mathematics: Philadelphia, PA.

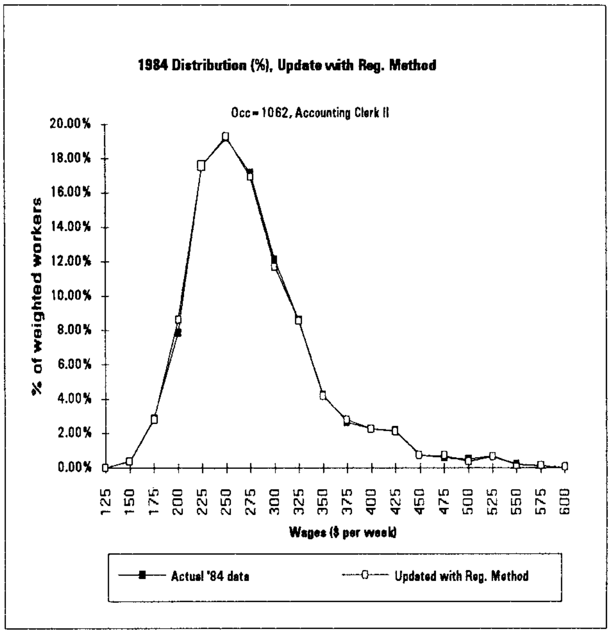
Occ. Code	Job Title	mddf
1061	Accounting Clerks I	0.009739
1141	Secretary I	0.003263
1144	Secretary IV	**0.007786**
2111	Computer System Analyst I	0.007189
2221	Computer Programmer I	0.012304
2223	Computer Programmer III	**0.060140**
4021	Guards I	0.013657
4030	Janitors	0.008907
4081	Truckdriver, light	0.015810
4082	Truckdriver, medium	0.009450

The two numbers in bold print correspond to the two graphs that were just displayed. The graphs that showed the most deviation from the actual data had a test statistic with a value greater than .01, such as the Computer Programmer job (.06). However, most occupations in the study were less than .01 and there was very little effect, if any, on the graphs, as we saw with the graph of the Secretary I wages (.007).

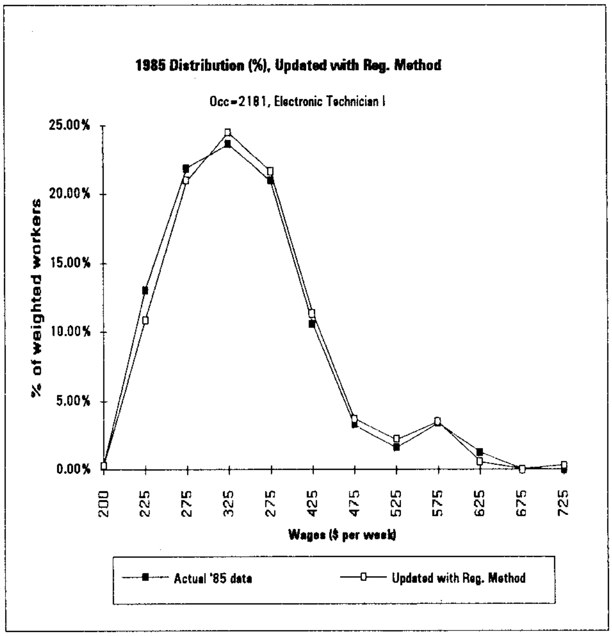
We also considered an alternate method of updating the wage data for the areas not surveyed in a given year. This method, referred to as the regression-based method of updating wage data, computes the rate of change in mean wages for each occupation from data common to earlier and later surveys. The rate of change, from the previous year to the current year, is calculated for each occupation and this factor is applied to the data in the same manner as the ECI factor (i.e., to the previous year's wage records for those areas that are not going to be surveyed). Instead of using an independent index such as the ECI to provide a factor by which to move the wage data, this method lets the survey data dictate the update factor that will be used.

The problems with this regression-based method of updating are that it requires more work to implement because it requires computing a rate of change for each occupation, instead of each occupational group, and it uses both certainty and non-certainty survey areas to update wage data for only non-certainty areas. Intuitively, the rate of change of just the surveyed non-certainty areas should more accurately reflect the changes in the wage data of the deselected non-certainty areas, but eliminating the certainty areas from the calculations would lead to insufficient data for some occupations since the certainty areas are the largest areas being surveyed.

As with the ECI method of updating, our tests showed that the distributions were relatively unaffected by the regression-based method of updating. Again, graphing the distributions show that the distributions displayed very little deviation between the actual data and the updated data. The graph below is the Accounting Clerks job and the two graphs are almost identical.



The second graph is the Electronics Technicians job which showed a slight deviation between the two graphs.



The greater effect on the graph of this distribution agrees with what was calculated by our test statistic.

Occ. Code	Job Title	mddf
1061	Accounting Clerks I	0.009482
1062	Accounting Clerks II	**0.006995**
1141	Secretary I	0.006804
2111	Computer System Analyst I	0.005832
2221	Computer Programmer I	0.011709
2181	Electronics Technician I	**0.031189**
4021	Guards I	0.008903
4030	Janitors	0.009699
4081	Truckdriver, light	0.011155
4082	Truckdriver, medium	0.008585

Again, this table shows that the Accounting Clerk job, whose graph was not affected by the update data, had a value of less than .01 (.006). And the Electronics Technician job, whose graph showed a little more deviation, has a value greater than .01 (.03). One other observation is that for the blue-collar occupations, such as Guards, Janitor, and Truck drivers, the value of the test statistic seemed to be closer to .01 or greater when compared to the other occupations in the study. This would suggest that the blue-collar jobs are more sensitive to the update process.

The update study showed that updating up to one half of the non-certainty areas' wage data had little effect on the distribution of workers across wages for most occupations. Nevertheless, there are differences among occupation groups in the effectiveness of the updating. The blue-collar jobs proved to be the most sensitive to the updating process but were still within an acceptable range for publication. A reason for this could be that wages for blue-collar occupations are based more on local labor markets so when pairing areas, one area may not accurately represent the other. Also, blue-collar occupations have a higher variance in general due to the union/non-union factor. Overall, there was no substantial difference between the results achieved using the ECI updating method or the regression-based method of updating.

For our purposes, we recommended using the ECI update method applied to the 1993 wage estimates, for the 1994 deliverable mainly because it would be easier to implement given the time constraints. Also we felt that the ECI method was better for the present time because the OCSP may be undergoing further changes which would make the ECI update method the clear choice for updating wage data.

For the blue-collar jobs, we found that the effect on the distributions was minimized by using the ECI index for the blue-collar major occupational group. For all jobs except blue-collar jobs, the White-Collar less Sales ECI index was satisfactory. Further study will be needed to assess the difference between the updating methods in terms of which is more efficient to use considering time and cost constraints, given that there was no sharp difference between the update methods where the reliability of the estimates is concerned. And the efficiency of each method will have to be reassessed as the Occupational Compensation Survey Program continues to evolve in an attempt to publish a greater amount of wage related data.