Application of Piecewise Quadratic Density Estimator to OES Wage Data

Teresa E. Hesley and Martha Duff, Bureau of Labor Statistics October 2009

Keywords: Wage Intervals, Density Estimator, Occupational Employment Statistics

Abstract

The Occupational Employment Statistics (OES) survey conducted by the U.S. Bureau of Labor Statistics (BLS) collects occupational wage data within pre-defined wage intervals. Current practice is for OES to use point data from the National Compensation Survey (NCS), also conducted by BLS, to derive interval means used in calculation of occupational wage estimates. In past analysis this method corrected the slight bias associated with using interval mid points or geometric means. This paper examines an alternative for estimating wages when employment has been reported in pre-defined wage intervals. We apply O'Malley's Piecewise Quadratic Density Estimator (PQDE) to OES wage interval data. This paper assesses the resulting OES wage estimates.

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

I. Introduction

The Occupational Employment Statistics (OES) Survey creates occupational employment, mean wage rate, and percentile wage rate estimates and associated sampling errors for detailed areas and industries. Data are collected from 1.2 million establishments in six semiannual panels over a three year period. OES collects wage rates of workers in 12 consecutive, non-overlapping wage intervals rather than individual wage rate point data in order to reduce response burden on employers and improve response rates. Interval widths are set so that the maximum relative error is approximately equal in each interval. The OES wage interval boundaries and widths as of the May 2008 estimation are provided in Table 1.

Table 1. OES Wage Intervals

OES Intervals	Lower Bound ($)	Upper Bound ($)	Interval Width ($)
A	5.15	7.49	2.35
B	7.50	9.49	2.00
C	9.50	11.99	2.50
D	12.00	15.24	3.25
E	15.25	19.24	4.00
F	19.25	24.49	5.25
G	24.50	30.99	6.50
H	31.00	39.24	8.25
I	39.25	49.74	10.50
J	49.75	63.24	13.50
K	63.25	79.99	16.75
L	80.00		∞

In an effort to produce the best estimates possible, OES has previously compared several techniques to estimate the average wage rate for workers in each interval. These included using the interval midpoints, and using geometric means. Prior research showed that for mean hourly wages, the arithmetic mean performed well, geometric mean performed better, and mean wages calculated from the National Compensation Survey (NCS) performed best.

OES currently uses national parameters derived from NCS point data for the 12 wage rate interval means and variances used in calculation of OES wage rate estimates. While the NCS collects individual wage data, it has a sample size of only 36,000 establishments. OES collects 1.2 million establishments to produce estimates for much more detailed area and industry domains than the NCS. Consequently, national interval means are applied to more detailed strata, which may not be well represented by the high level means. A method that utilizes the large quantity of interval wage data collected by OES to produce wage estimates would be desirable.

O'Malley (2008) has developed a simple density estimator that smoothes histogram interval data into a piecewise quadratic function. The estimator has the advantage of being able to more fully use the information available in large quantities of interval data by considering both the proportions in the intervals and their relationship to adjacent intervals. Successful application of such an estimator would eliminate the need to use an outside data source for interval means and may provide more reliable data at greater levels of geographic and occupational detail. This paper presents research conducted in applying O'Malley's piecewise quadratic density estimator to OES data.

II. Piecewise Quadratic Density Estimation[1]

A. General Description

In PQDE, employment within a wage interval is represented by the area under a curve drawn to show the estimated relationship between employment and hourly wage rate within the wage interval. PQDE is founded on two guiding principles:

- The area in each interval of a frequency histogram is preserved.
- The curve should be somewhat smooth with no large spikes or jumps between intervals

These two principles allow for a simple, intuitive, closed form density estimator which seems well suited to the large samples common to OES.

For OES data, we first create a histogram of wage intervals where the height of the histogram in each interval represents the employment in that wage interval. We then use the area in the histogram and the interval wage value boundaries to derive a quadratic equation to represent the relationship between employment(y) and wages(x) in this interval. The equation is of the form:

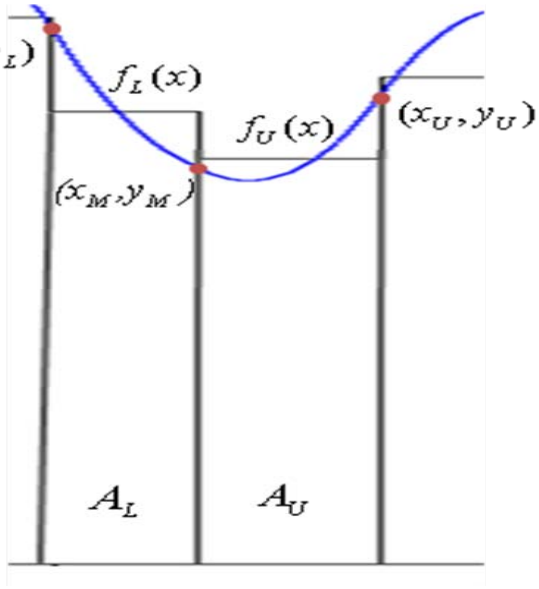
$$y = f(x) = a_2x^2 + a_1x + a_0$$

Note that this procedure applies only to center intervals where adjacent histogram bars are present on the right and left. The y-values in the following equations are initialized at

[1] For a more complete description of PQDE see O'Malley, Meghan. "Density Estimation for Interval-Censored Economic Data." Section on Survey Research Methods – JSM 2008, and references therein.

the average of the heights of the adjacent histogram bars. To maintain continuity, the y-values are set to zero if either adjacent interval is zero.

Illustration with polynomial derivations:



The constraints for the lower polynomial, $f_L(x)$, are as follow:

$$A_L = \int_{x_L}^{x_M} f_L(x)dx = a_{L,2}\frac{x_M^3 - x_L^3}{3} + a_{L,1}\frac{x_M^2 - x_L^2}{2} + a_{L,0}(x_M - x_L)$$

$$y_L = f_L(x_L) = a_{L,2}x_L^2 + a_{L,1}x_L + a_{L,0}$$

$$y_M = f_L(x_M) = a_{L,2}x_M^2 + a_{L,1}x_M + a_{L,0}$$

Note: A_L is the area under the curve between x_L and x_M.

We can solve for the coefficients:

$$a_{L,2} = 3\frac{y_L + y_M}{(x_L - x_M)^2} - 6\frac{A_L}{(x_L - x_M)^3}$$

$$a_{L,1} = \frac{y_L - y_M}{x_L - x_M} - (x_L + x_M)a_{L,2}$$

$$a_{L,0} = y_M - x_M\frac{y_L - y_M}{x_L - x_M} + (x_L x_M)a_{L,2}$$

And, similarly, the constraints for the upper polynomial, $f_U(x)$,can be written as:

$$A_U = \int_{x_M}^{x_U} f_U(x)dx = a_{U,2}\frac{x_U^3 - x_M^3}{3} + a_{M,1}\frac{x_U^2 - x_M^2}{2} + a_{U,0}(x_U - x_M)$$

$$y_U = f_U(x_U) = a_{U,2}x_U^2 + a_{U,1}x_U + a_{U,0}$$

$$y_M = f_U(x_M) = a_{U,2}x_M^2 + a_{U,1}x_M + a_{U,0}$$

We can solve for the coefficients:

$$a_{U,2} = 3\frac{y_U + y_M}{(x_U - x_M)^2} - 6\frac{A_U}{(x_U - x_M)^3}$$

$$a_{U,1} = \frac{y_U - y_M}{x_U - x_M} - (x_U + x_M)a_{U,2}$$

$$a_{U,0} = y_M - x_M \frac{y_U - y_M}{x_U - x_M} + (x_U x_M)a_{U,2}$$

Polynomials derived based on the midpoints and area in each interval may not meet our goal of smoothness, so the connector points are algebraically adjusted to reduce differences between slopes of adjacent polynomials. The goal is to choose y_M to minimize the difference between the slopes of the lower and upper polynomials. This is an iterative process, though rarely are improvements seen after the second iteration. This adjustment is performed as follows:

Let $\left| f_L^{'}(x_M) - f_U^{'}(x_M) \right| = 0$

$$\left|(2a_{L,2}x_M + a_{L,1}) - (2a_{U,2}x_M + a_{U,1})\right| = 0$$

$$y_M = \frac{3\left(\frac{A_L}{(x_M - x_L)^2} + \frac{A_U}{(x_U - x_M)^2}\right) - \left(\frac{y_L}{x_M - x_L} + \frac{y_U}{x_U - x_M}\right)}{2\left(\frac{1}{x_M - x_L} + \frac{1}{x_U - x_M}\right)}$$

$$y_M = \max\{y_M, 0\}.$$

The mean of any interval is calculated as the area inside the interval divided by the width of the interval:

$$MeanWage_{ML} = \frac{A_L = \int_{x_L}^{x_M} f_L(x)dx = a_{L,2}\frac{x_M^3 - x_L^3}{3} + a_{L,1}\frac{x_M^2 - x_L^2}{2} + a_{L,0}(x_M - x_L)}{(x_M - x_L)}$$

B. Handling Negative Regions

It is possible for the parabolas in certain intervals to go negative. This can happen when an interval with a small proportion of employment is between intervals with large proportions of employment or when a small proportion of employment is adjacent to a very large proportion of employment. For example, within a specific occupation entry level workers may receive an Interval B wage rate that generally jumps directly to an Interval D wage rate when they achieve journeyman status. In this case, the parabola for Interval C, where there is little employment, might dip below zero. When a negative region occurs, the parabola should be replaced by one or more lines which drop from the larger interval to zero in a way that preserves area.

C. Handling End Intervals

In OES, the beginning and ending wage intervals are different from central intervals and require special handling.

Interval A is bound on the left hand side, at the federal minimum wage. For this interval, the only constraints are the area under the histogram and the connector point on the right hand side of the interval. We represent this area with a linear polynomial. We may need to adjust this interval to include natural spikes in the data due to clumping of employment at minimum wage.

Interval L is unbounded on the right hand side. The location and spread of the data in this interval can have a large impact on the mean and higher percentiles. In order to make estimates from this type of censored data without relying on strong assumptions or an outside source, the extent of this problem must be limited by raising the lower bound of the last interval until it is large enough that only a small tail remains in the rightmost interval. An exponential distribution can then be used to describe the tail. We analyze this situation further in the next section.

III. Application of Piecewise Quadratic Density Estimator to OES Data

Our goal was to incorporate the PQDE described in section II into our current mean wage estimation system and analyze the resulting estimates. Modifications to various aspects of our system were required. Most notable were changes needed in handling

A. Wage Interval L, our upper unbounded wage interval, and
B. Variances on mean wage estimates.

We discuss each of these in greater detail below.

A. Interval L

Interval L, the highest wage interval on the OES survey form (Attachment A.) is used to record employees making 80 dollars per hour and above[2]. This interval only contains a small portion (roughly 2%) of OES data. Interval L presents special estimation problems because it is open-ended. The quadratic polynomials used to represent data in the central bounded OES wage intervals cannot be applied to Interval L which we instead represent with an exponential function. For the exponential function to accurately represent this data it is important that the data in this interval form a small tail. Because the exponential function assumes decay, it is particularly important that there be no mode in this data. About 30 OES occupations have 10% or more of their employment in Interval L. A specific example is anesthesiologists, which report over 75% of their employment in interval L. Since the exponential distribution used to describe this interval will not work well with large amounts of data or when the data contains a mode, occupations with larger employment levels in L must have the bulk of that employment redistributed into a bounded interval where PQDE can be applied.

We examined two options for redistributing data from the open ended L into a bounded interval.

- Extend the lower bound of Interval L (upper bound of K) out further. This option moves the interval bound up so that the mode is now in a larger Interval K and Interval L has become smaller.

[2] OES wages are cut-off at $480 per hour which is approximately equal to one million dollars per year, assuming a 40 hour work week. This extremely high upper bound is treated as an unbounded interval for estimation purposes.

Or

- Create a new interval (Interval M) for the purposes of estimation only. Divide employment originally in L between L and M with M now being the open ended interval and L, the bonded interval containing the portion of employment with the mode.

We found the creation of a new Interval M more favorable than extending the boundary between K and L. Simply moving up the lower bound of L caused Interval K to be quite large in some cases.

Following is the method we used to determine the location of the new Interval M lower bound (Interval L upper bound) for occupations with a mode in Interval L.

1. Using the NCS data[3], create a dataset that contains all observations in NCS with wages above 80 but below 480 dollars (the million dollar a year cutoff). Also, exclude pilots (soc = 532011) from this dataset. (Pilots have unique work hour/pay schedules that don't mix well with other occupations when trying to develop a universal estimation system.)

2. Calculate the median and 66th percentile for all NCS observations in L..

3. Determine NCS occupations with at least five observations in interval L, and for each of these occupations calculate the mean wage in L.

4. Classify each occupation into one of three categories:

   Part A. Occupations whose L mean is below the overall Interval L median calculated in step 2.
   Part B. Occupations whose Interval L mean is above the overall Interval L median.
   Part C. Occupations with less than five NCS observations in Interval L or whose Interval L mean is equal to the overall Interval L median.

   For recent NCS data, the distribution of occupations into these parts was as follows:

Table 2. Distribution of Occupations into Interval L Categories

Occupation Group	Number of Occupations	Percentage of Occupations
Part A	37	4.6
Part B	49	6.1
Part C	715	89.3
All	801	100.0

5. Calculate a 2/3rds percentile for Part A; this 2/3rds percentile will become the upper bound of L. Part B is handled similarly. For Part C, the overall Interval L

[3] This method does reintroduce the use of data from an outside source into our estimation procedure, but only to set wage interval boundaries and not for calculation of means.

2/3^{rds} percentile calculated in step 2 will become the upper bound of L. A new interval, (Interval M), will be created as an open ended interval containing the remaining upper third percentile for Parts A, B, and C. Charts 1, 2, and 3 illustrate selection of the new bounds.

Chart 1. New Interval Bound: Part A

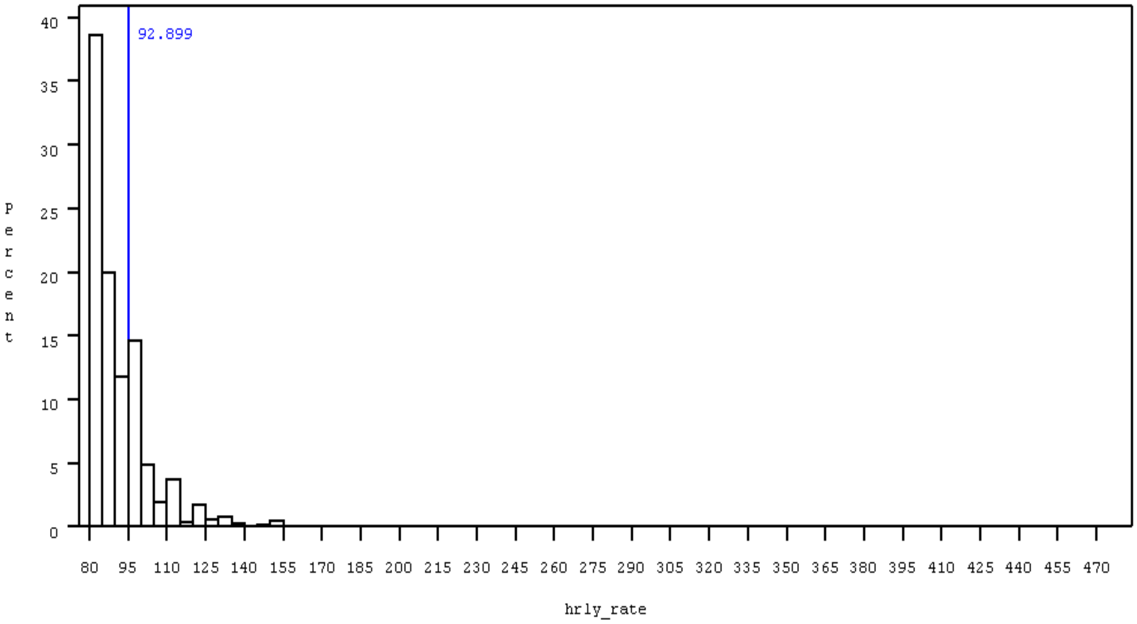


Chart 2. New Interval Bound: Part B

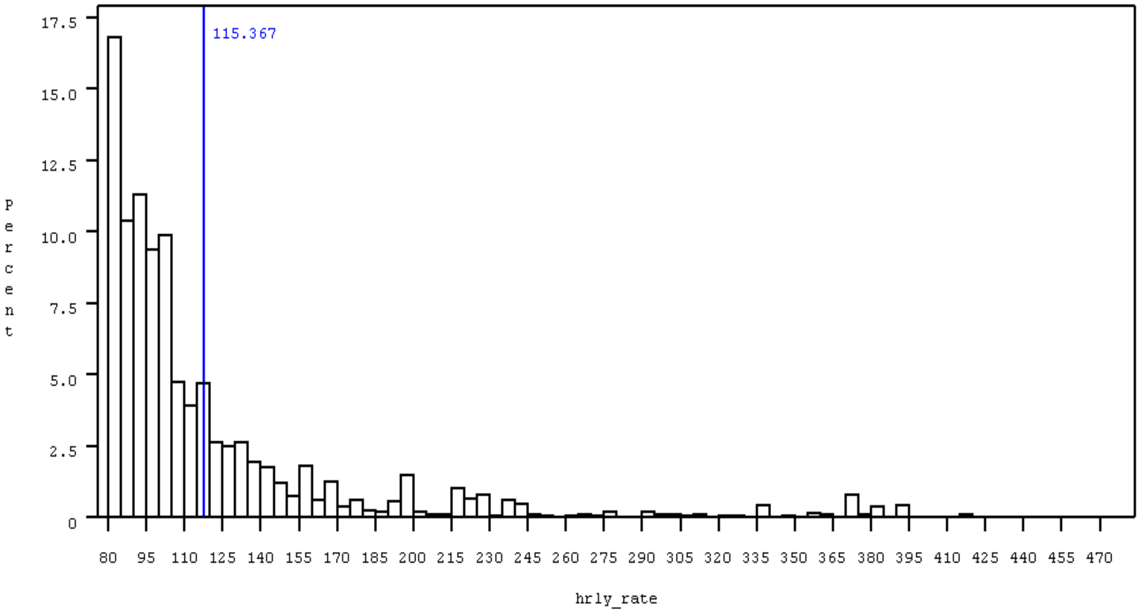
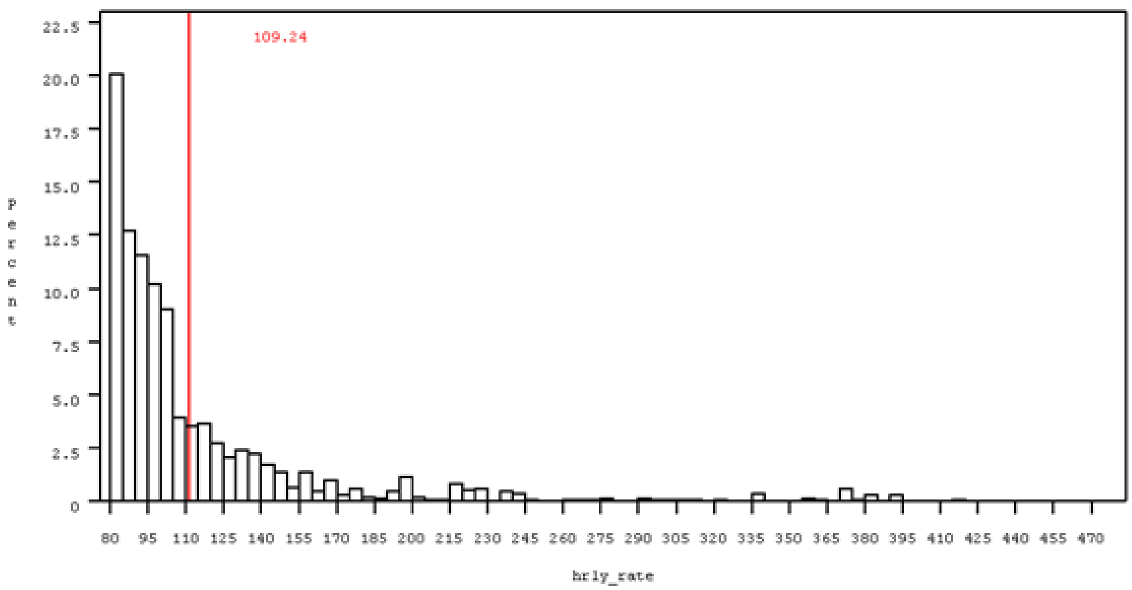


Chart 3. New Interval Bound: Part C



6. In the event an occupation is found to have a mode in Interval L, we redistribute the employment that was originally in L for individual occupations based on bounds determined above for their appropriate occupational grouping (Part A, B, or C) as follows:

$$L_{Emp} = 0.66 * L_{OrigEmp}$$

$$M_{Emp} = 0.33 * L_{OrigEmp}$$

Where: L_{Emp} - Newly distributed employment for interval L

$L_{OrigEmp}$ - Original employment for interval L

M_{Emp} - Newly distributed employment for interval M

B. Variance Estimates

OES wage variance estimates currently use variance components calculated using NCS point data. Since our PQDE method does not use NCS data in the direct estimation of interval means, a new variance estimator must be put into place. We developed and implemented a random group Jackknife variance estimator (Wolter, 1985) for OES wages similar to the estimator now used in OES employment estimation. In order to show the effectiveness of the PQDE in comparison with our current estimation method, we applied our new Jackknife variance estimator to both the PQDE derived estimates and the OES one panel estimates.

Jackknife Variance Estimator:

1. Calculate Overall Estimate

2. Divide estimation data file into six Random Groups. Five groups of non-certainty establishments and one group of certainty establishments.

3. Remove one of the non-certainty Random Groups (group numbers 1 through 5) and adjust benchmark weight of non-certainty units by taking the benchmark employment of the total and dividing it by the benchmark employment with the group removed. The weights of the certainty units (random group number 6) remain intact.

4. Recalculate the estimate.

5. Calculate pseudo-estimates as follows:

$$\hat{\theta}_\alpha = 5\hat{\theta} - (5-1)\hat{\theta}_{(\alpha)}$$

Where:

$\hat{\theta}_\alpha$ - pseudo-estimate

$\hat{\theta}$ - Overall Estimate

$\hat{\theta}_{(\alpha)}$ - estimate with Random Group removed

6. Obtain the average of the pseudo values

$$\bar{\hat{\theta}} = \sum_{\alpha=1}^{5} \frac{\hat{\theta}_\alpha}{5}$$

7. Estimate the variance

$$v\left(\bar{\hat{\theta}}\right) = \frac{1}{5(5-1)} \sum_{\alpha=1}^{5} \left(\hat{\theta}_\alpha - \bar{\hat{\theta}}\right)^2$$

8. Calculate Standard Error as the square root of the variance

9. Calculate the Relative Standard Error by dividing the SE by the overall estimate. Multiply by 100 to express as a percentage.

IV. Results

A. Output Overview

National level estimates of mean wages and associated variances were created for 801 occupations using the most recent panel (May 2008) of OES data. One panel of OES data includes interval wage data from approximately 200,000 establishments. Following are highlights of our results[4]:

- The average difference between PQDE and current OES mean wages across all occupations was 0 cents per hour.
- The average absolute difference in mean wages between the two methods was 16 cents per hour with 307 occupations showing an increase in estimated mean hourly wage while 494 occupations showed a decrease in estimated mean hourly wage.
- The average percent change in mean wages was 0.59%.
- Only 14.36% of occupations saw a change in mean wage of more than 1%.
- Less than 1% of occupations (7 of 801) saw a change in mean wage greater than 5%.
- Dentists had the largest dollar wage change increasing from $75.35 in OES to $81.08 in PQDE, a difference of $5.73. (Over 25% of dentists were reported in Interval L.)
- The largest percentage change in wages occurred for Medical Appliance Technicians whose wages decreased 8.1% from $19.76 in OES to $18.17 in PQDE.

We grouped occupations by broad mean wage categories to see if there were any differences in the comparison of PQDE estimates to OES estimates between higher and lower wage earning occupations. The choice of range was non-scientific. All differences were small, but higher wage earning occupations tended to see increases in mean wage while lower earning occupations saw wage decreases.

Table 3. Mean Wage Comparison OES to PQDE by Wage Categories

Occupations		Averages			
Mean Wages (OES)	Number	Absolute % Change	Actual % Change	Absolute $ Change	Actual $ Change
$40.00 and over	89	1.12%	0.19%	$0.67	$0.17
$20.00 - < 40.00	320	0.40%	0.01%	$0.11	$0.00
Under $20.00	392	0.62%	-0.33%	$0.08	-$0.05
All	801	0.59%	0.14%	$0.16	$0.00

We expected to see the increased mean wage in higher earning occupations, because of changes made to Interval L to accommodate the PQDE. The redistribution of L into two Intervals (L and M) when modes were found in L, allows higher wage earning occupations to have a higher mean wage assigned via Interval M.

We don’t yet have an explanation for the tendency toward slightly lower mean wages in the lower earning occupations. Possibly the non PQDE derived Interval A is a factor. In the following section we discuss Interval A in more depth.

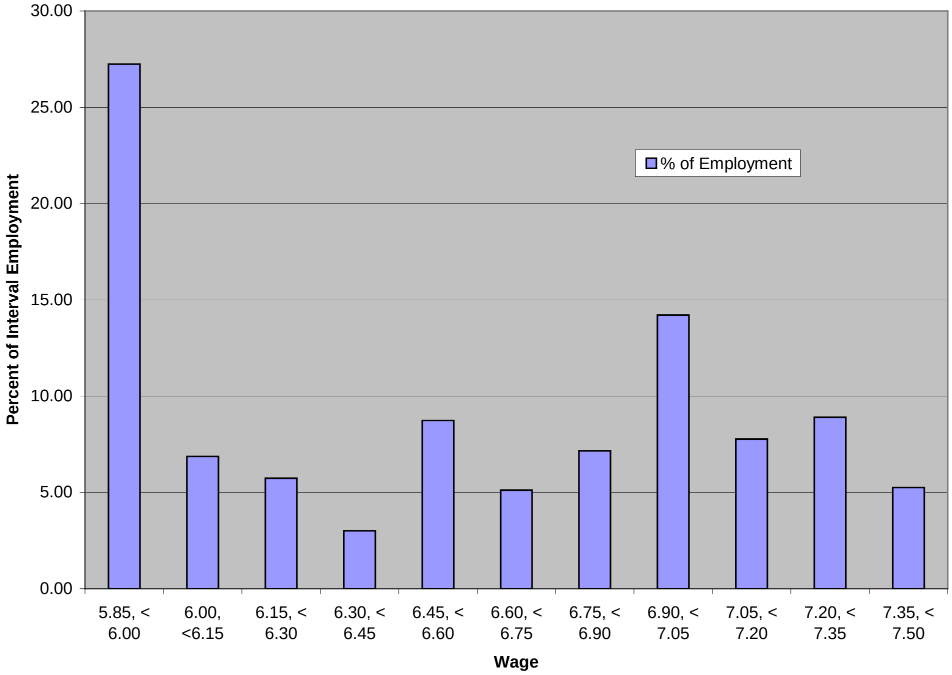
B. Interval A

Density estimation allows us to make use of additional information contained in the relationships between intervals. Because Interval A has no interval before it from which to gather information, it is not suited for PQDE. We began with a straight line estimator

[4] These results are based on one panel of data only and are not official OES mean wage estimates.

for Interval A as derived by O'Malley. When we took a closer look at Interval A, it became apparent that a simple straight line is not appropriate for Interval A. The Chart below shows the distribution of NCS nationwide employment in Interval A as a percentage of the total Interval A employment. Note the large spike at the national minimum wage ($5.85).

Chart 4 Wage Interval A Employment Distribution as a Percent of the Total, Nationwide



A more reasonable approach would be to distribute employment within Interval A including spikes where we find spikes in the NCS data for this wage range. Specifically, we found employment spikes at the minimum wage itself, and at various points throughout the interval, most of which represent state minimum wages that exceed the federal minimum wage. Table 4 shows individual wage points with employment spikes equal or greater than 5 percent of the total employment in Interval A based on the national NCS sample.

Table 4. Interval A Employment Spikes

Wage	% of Interval A Total
5.85	26.23
7.00	10.86
6.50	6.64
6.00	5.29

Work is underway to adjust Interval A to include one or more wage spikes when distributing employment.

C. Variance

We calculated variance estimates on mean wage estimates based on one panel of data for national occupational groupings using current OES methodology and PQDE using our random group Jackknife method. The PQDE produced mean wage estimates with a lower variance for 62% of occupations. The average change in mean wage RSE from OES to PQDE was a decrease of 0.19. The average percent change in mean wage RSE from OES to PQDE was -9.74%.

V. Conclusion

The Piecewise Quadratic Density Estimation method shows some promising results. The central bounded OES wage Intervals (B through K) show results indicating that they are adequately represented by the PQDE at the national major occupation group level of detail. Our procedure to eliminate modes in the upper open-ended Interval L, allows this interval to be reasonably estimated using an exponential function. The lowest wage interval (Interval A) requires further research.

VI. Future Research

A. Aging Wages from Prior Panels

OES collects its data semi-annually over three years. To make the wages 'current', OES calculates aging factors for older data panels from the Employment Cost Index (ECI) survey. The ECI survey measures the rate of change in compensation for ten major occupation groups on a quarterly basis. These aging factors are used to update OES wages, by occupation group, from the previous five panels to current levels. Our current estimation methodology applies these updates to NCS derived wage interval means. In order to apply O'Malley's PQDE it will be necessary for us to adapt our current aging procedure to the new estimator.

B. Small Domains

The OES survey produces a large number of very detailed estimates, and so the quality of the PQDE in small domains is of particular interest. We will be examining the performance of the PQDE in detailed industry and area cells, where in some cases there will be few observations.

C. Estimates of Percentiles

Based on previous research at BLS, we know that our current methodology produces biased estimates of percentiles. We plan to take a closer look at the percentile estimates produced by the PQDE method.

D. Incorporating Point Data

OES now receives some point data submitted electronically from various business establishments in the survey. Currently these data are not used in point data form. It would be advantageous to use point data, where available, along with interval data where point data is not available. A procedure to use a combination of point and interval data would allow OES to make the best estimates possible.

E. Interval A

Additional work is needed to more accurately distribute employment within OES Interval A. The distribution should recognize spikes at national and state minimum wage points where research shows natural clumping of data points.

References

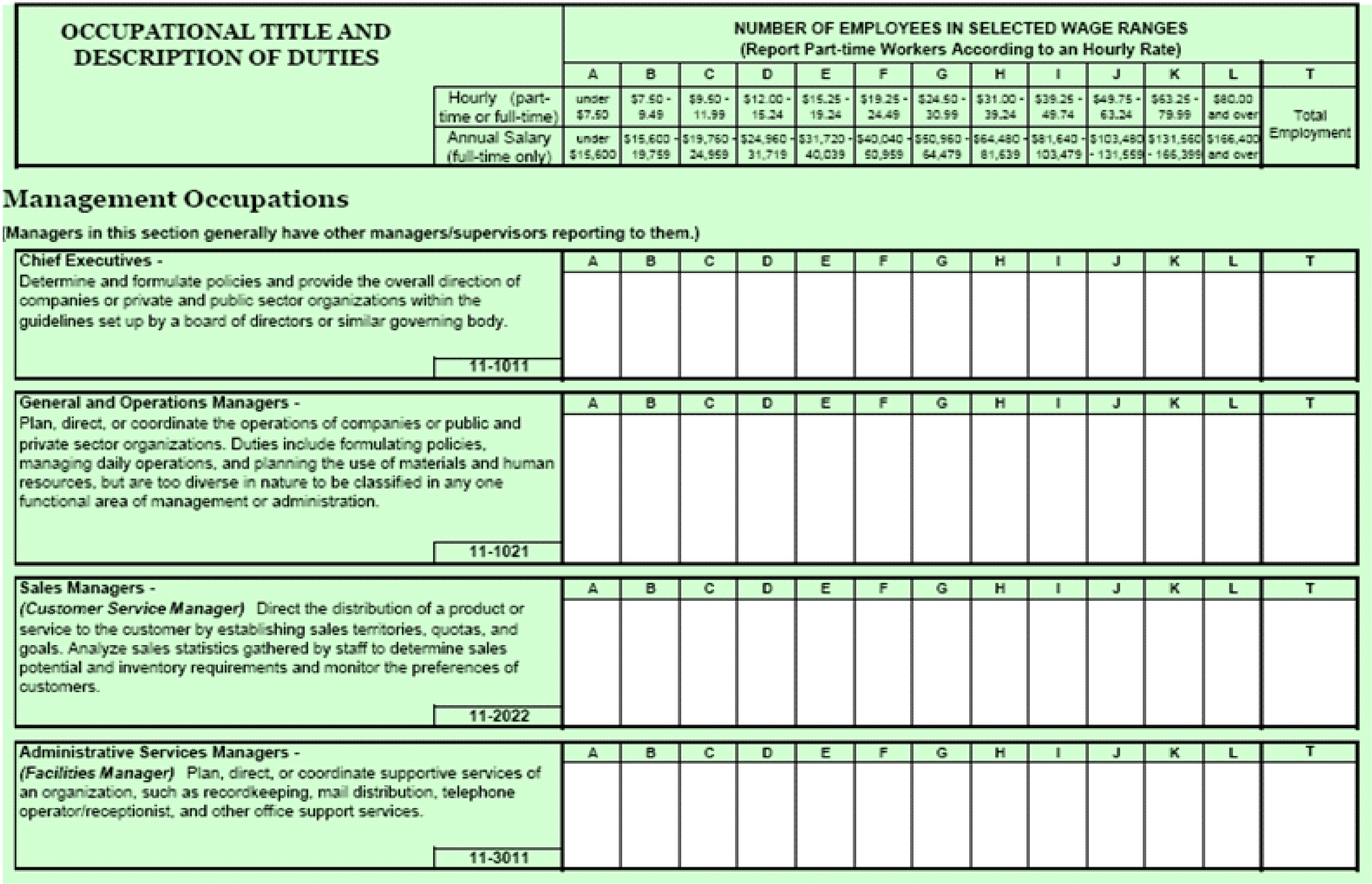
O'Malley, Meghan. 2008. "Density Estimation for Interval-Censored Economic Data". In JSM Proceedings Section on Survey Research Methods. 1204 – 1211.

Wolter, Kirk M. 1985. *Introduction to Variance Estimation* NewYork: Springer Verlag.

"Survey Methods and Reliability Statement for the May 2008 Occupational Employment Statistics Survey", http://www.bls.gov/oes/current/methods_statement.pdf

Silverman, B.W 1986. *Density Estimation for Statistics and Data Analysis* Washington, DC: Chapman & Hall/CRC

Attachment A: Example of OES Survey Form for NAICS 213000 (Support Activities for Mining)



OCCUPATIONAL TITLE AND DESCRIPTION OF DUTIES		NUMBER OF EMPLOYEES IN SELECTED WAGE RANGES (Report Part-time Workers According to an Hourly Rate)												
		A	B	C	D	E	F	G	H	I	J	K	L	T
	Hourly (part-time or full-time)	under $7.50	$7.50 - 9.49	$9.50 - 11.99	$12.00 - 15.24	$15.25 - 19.24	$19.25 - 24.49	$24.50 - 30.99	$31.00 - 39.24	$39.25 - 49.74	$49.75 - 63.24	$63.25 - 79.99	$80.00 and over	Total Employment
	Annual Salary (full-time only)	under $15,600	$15,600 - 19,759	$19,760 - 24,959	$24,960 - 31,719	$31,720 - 40,039	$40,040 - 50,959	$50,960 - 64,479	$64,480 - 81,639	$81,640 - 103,479	$103,480 - 131,559	$131,560 - 166,399	$166,400 and over	

Management Occupations

(Managers in this section generally have other managers/supervisors reporting to them.)

	A	B	C	D	E	F	G	H	I	J	K	L	T
Chief Executives - Determine and formulate policies and provide the overall direction of companies or private and public sector organizations within the guidelines set up by a board of directors or similar governing body. **11-1011**													

	A	B	C	D	E	F	G	H	I	J	K	L	T
General and Operations Managers - Plan, direct, or coordinate the operations of companies or public and private sector organizations. Duties include formulating policies, managing daily operations, and planning the use of materials and human resources, but are too diverse in nature to be classified in any one functional area of management or administration. **11-1021**													

	A	B	C	D	E	F	G	H	I	J	K	L	T
Sales Managers - *(Customer Service Manager)* Direct the distribution of a product or service to the customer by establishing sales territories, quotas, and goals. Analyze sales statistics gathered by staff to determine sales potential and inventory requirements and monitor the preferences of customers. **11-2022**													

	A	B	C	D	E	F	G	H	I	J	K	L	T
Administrative Services Managers - *(Facilities Manager)* Plan, direct, or coordinate supportive services of an organization, such as recordkeeping, mail distribution, telephone operator/receptionist, and other office support services. **11-3011**													