Empirical Evaluation of Imputation Methods on Quarterly Census of Employment and Wages (QCEW) Data

Marek W. Kaminski and Vinod Kapani
Bureau of Labor Statistics, Washington, DC

Abstract

The U.S. Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW) program currently uses each establishment's year-ago trend in imputing missing employment and wages data. Ratio method is introduced which is using current trend of employment and wages. An empirical evaluation of well established methods, namely ratio and nearest–neighbor, is undertaken. This paper presents the analytical evaluation of these methods using current trends in the data. The reported data is simulated for imputing employment and wages on a random sample from QCEW Longitudinal Database (LDB). Both methods utilize exclusion criterion for removing influential observations from the computations. Finally, we offer comparisons of both methods, at stratum and aggregate levels, percentage relative errors.

Key words: Employment imputation, wages imputation, simulation of missing data, ratio method, nearest neighbor method, percent relative error, influential observations

Disclaimer

Views expressed in this document are those of the authors and do not necessarily reflect the views or policies of the Bureau of Labor Statistics.

1. Introduction

1.1 Data: Quarterly Census of Employment and Wages (QCEW) earlier known as ES 202 program is a quarterly census of all U.S. establishments that are subject to unemployment insurance taxes. The census is conducted by each state and the data is then sent to the central BLS office in Washington DC. Each quarter establishments report total employment for each month of the quarter and quarterly total wages. Each establishment is represented as one record. A part of a typical record in QCEW as stored in the Longitudinal Database might look as follows:

ldb_num	yr_qtr	state_fips	naics	m1_empl	m2_empl	m3_empl	tot_wage
11…	20014	06	311421	98	105	101	1045612

Fig. 1 Example of a record in QCEW (we only show fields that are referenced in this article)

where lbd_num is unique number assigned to every establishment, yr_qtr is year and quarter of the record (in this example it is the fourth quarter of year 2001), state_fips is a two digit number of state where establishment is located, naics is a North American Industry Classification System, it is a 6 digit number which identifies type of industry to which establishment belongs to (in this example, 311421 stands for Flavoring Syrup and Concentrate Manufacturing), m1_empl, m2_empl, and m3_empl are employment for the first, second and third month of the quarter and tot_wage are total wages paid to employees during the whole quarter.

Quarterly total wages are used to determine unemployment taxes. QCEW longitudinal data base contains records of approximately 8.5 million of establishments per quarter, with approximately 130 million of employment in a month of the quarter. QCEW data is used as a sampling frame for BLS conducted surveys including Current Employment Statistics (CES), and Job Opening and Labor Turnover Survey (JOLTS). These surveys produce principal federal economic indicators. QCEW is also used for "benchmarking" estimates obtained from these programs (benchmarking is an adjustment of estimates from a program to match QCEW counts). Business Employment Dynamics (BED) produces statistics on labor market based on QCEW data. QCEW employment and wages data are published each quarter for nation, states, metropolitan areas, and counties. A number of private and public sector activities are based on QCEW data. It is also used as a major component of national and state personal income statistics and gross domestic product (GDP). Thus everything that exists in QCEW data, and is performed on QCEW data, indirectly affects other important programs. Imputed employment and wages are a part of QCEW data. Every quarter BLS performs imputation of data for about 300,000 establishments, imputing about 3,300,000 of employment, and about 27 billions of dollars of wages. Therefore, research on imputation methods in QCEW has great importance.

1.2 Present method of imputation: Currently, QCEW is using year ago trend of the individual establishment for estimation of missing employment and wages. For example if an establishment fail to report employment in June 2009, and wages for the second quarter of 2009, while one year ago nonzero data of employment and wages is available, then the following formulas could be used for estimation of employment and wages. The employment is reported for July and August. The present example is a large simplification of the actual process:

$$(1) \qquad \widehat{empl}_{June\_2009} = \frac{empl_{June\_2008}}{empl_{May\_2008}} \times empl_{May\_2009}$$

(2) $$\widehat{wages}_{2nd\_quarter\_2009} = \frac{wages_{2nd\_quarter\_2008}}{wages_{1st\_quarter\_2008}} \times wages_{1st\_quarte\_2009}$$

Using exactly one year ago trend has some disadvantages. It is incompatible to other BLS programs, which use *current trend* data for estimations. It also does not reflect current economic changes.

1.3 Proposed methods of imputations: In this article the two methods are discussed namely ratio method, and nearest neighbor method use a current trend data. The *current trend* is understood as a change within data from no further than the past quarter, to the present time. *Current trend* is determined by ratio method and nearest neighbor method in different ways.

In ratio method, we find *current trend* of an entire **cell/stratum**. For employment, it is done by computing a ratio of total employment of all establishments in a cell from one month to the next month. For wages, it is done by computing a ratio of total wages of all establishments in a cell from a previous quarter to a current quarter. Such computed ratios measure *current trend* on a cell level.

In nearest neighbor method within a cell of establishments which requires imputation we find one establishment which we consider to be the nearest in the previous quarter. We call this nearest establishment *nearest neighbor*. Then compute the *nearest neighbor* employment ratios from month to month, and wages ratio from quarter to quarter. These ratios measure *current trend* of the *nearest neighbor*.

Both ratio method and nearest neighbor method use only the establishment's last reported values of employment and wages, together with estimated *current trend*, to estimate missing employment and wages.

2. Outline of Methodology

2.1 Basic definitions: In our studies, we use two consecutive quarters of data. The first quarter of data we call *previous quarter* and the second quarter of data we call *current quarter*. That gives us for each establishment 6 months of employment data and 2 quarters of total wages data. Establishments are divided into cells. Cells are sets of units that are homogeneous with respect to state, type of industry, and size of establishment during the third month of the previous quarter.

2.2 Sample selection for simulation: For a given two consecutive quarters a random sample is drawn for imputations. In a cell we count a number of establishments which did not report data during a current quarter. Then from the same cell we select a simple random sample of the same number of establishments which reported both employment and wages for both quarters. Thus, the selected sample in its distribution resembles the distribution of units which did not report in a current quarter.

2.3 Simulation of missing data, imputation, and evaluation of results: From a selected sample, we remove reported employment and wages in the current quarter and store this data for a future use. We treat this sample as if it were units that require imputation in the current quarter for employment and wages data. We perform imputations by ratio and nearest neighbor methods. Imputed values are compared to the stored actual values. If imputed values are far from actual values then we have large error, if imputed values are close to actual values then we have small error. In order to measure error, we compute percent relative errors between imputed and actual values on states level, national level by size class, and aggregate national level. Formulae are defined and demonstrated with examples in later sections.

For every two consecutive quarters from Jan. 2000 to June 2007 we perform imputations on a sample from *previous quarter* to *current quarter*, i.e. the last reported value of an establishment in the previous quarter together with observed trend is used for estimations of employment and wages in the current quarter. Imputations are performed for 90 months of employment and 30 quarters of wages.

3. Definition of Nearest Neighbor and Ratio Methods

The successful application of nearest neighbor and ratio methods solely depends on definition of cell structures and the exclusion of influential observations from computation for both methods. Cells should be as large as possible, to be as homogeneous as possible, and at the same time the cell structure have to be simple to make each method's application easy. These three constraints and the balance between them is the major challenge of this research.

3.1 Cells Definition: Cells are defined by state, 1 or two first digits of North American Industry Classification System (NAICS), size class and private ownership. The definition of size classes is based on the third month employment of the previous quarter. We define 7 size classes as follows: class0: employment 0, class1: employment 1 - 4, class2: employment 5 - 9, class3: employment 10 - 19, class4: employment 20 - 49, class5: employment 50 - 99, class6: employment 100 +. The class 0 is excluded from this paper as this group is imputed differently.

3.2 Influential observations: Influential observations take on extreme values; their inclusion or exclusion greatly influences the estimates. These observations generally occur in most surveys and censuses with low frequency. It is important to note that influential observations can be data representing real events that actually occurred or an error due to one of many reasons like response error and data entry error. Thus, one would expect influential observations to occur with low frequency on QCEW with over eight million records on a quarterly basis.

Both nearest neighbor and ratio method use cells for defining homogeneous sets of units. Both methods perform poorly if influential observations are part of cells. Removing a small percentage of influential observations from cells assures greater homogeneity without shrinking cell.

3.3 Definition of influential observations: Influential observations are a subset of the reported positive units. In order to determine establishments deemed as influential observations, we need to compute month-to-month ratios of employment as well as the ratio of wages between quarters for each reported establishment. The month–to–month employment ratio for an establishment is the ratio of the current month employment to the previous month employment. The quarter-to-quarter wage ratio is the ratio of current quarter wages to the previous quarter wages. By formulas:

$$(1) \qquad re_t = \frac{empl_t}{empl_{t-1}}, \quad t = 4, 5, 6, \quad \text{and} \quad rw = \frac{cw}{pw}$$

where: $empl_t$ is reported employment for month t, re_t is ratio of month employment for month t from month, t-1, cw is current quarter reported wages, pw is previous quarter reported wages, rw is ratio of wages. The table below summarizes the cut off ratio values for influential observations.

Table 1 Cut off values for influential observations by size classes.

Size Class	re_4	re_5	re_6	rw
(1) 1 - 4	5	5	5	9
(2) 5 - 9	5	5	5	6
(3) 10 - 19	5	5	5	6
(4) 20 - 49	5	5	5	5
(5) 50 - 99	5	5	5	4
(6) 100 +	5	5	5	3

The brief report on the research that led to defining cut off values is presented in section 5 of this paper. Before defining ratio and nearest neighbor methods we introduce terminology and notation which will be useful.

3.3.1 Terminology and Notation: In two consecutive quarters of data, previous quarter and current quarter, months are numbered from 1 to 6. By ***t*** we denote t^{th} month in the sequence of these two quarters, **c –** the current quarter, **p -** the quarter prior to the current quarter, **h -** cell, **do -** donor, **re -** recipient.

3.3.2 Positive Units are establishments with positive employment in the third month of the previous quarter and positive wages in the previous quarter.

3.3.3Reported Positive Units (Rep & Pos) are establishments reporting positive employment (>0) for the third month of the previous quarter and positive wages for the previous quarter.

3.4 Ratio Method: We compute monthly employment and quarterly wages ratios

$$(2)\quad RE_{h(t)} = \frac{\sum_{h(t)\cap(\text{Rep}\&\text{Pos - inf\_obser.})} employment}{\sum_{h(t-1)\cap(\text{Rep}\&\text{Pos - inf\_obser.})} employment}, \qquad t = 4,5,6$$

$$RW = \frac{\sum_{h(current)\cap(\text{Rep}\&\text{Pos - inf\_obser.})} TotalWages}{\sum_{h(previous)\cap(\text{Rep}\&\text{Pos - inf\_obser.})} TotalWages}$$

where $h(t) \cap (\text{Rep \& Pos - inf\_obser.})$ - set of establishments in cell *h* during the *t*-th month which are reported positive units, and are not influential observations,
$h(previous) \cap (\text{Rep \& Pos - inf\_obser.})$ - set of establishments in cell *h* during the previous quarter which are reported positive units, and are not influential observations,
$h(current) \cap (\text{Rep \& Pos - inf\_obser.})$ - set of establishments from the cell *h* during the current quarter which are reported positive units, and are not influential observations.

We use both type of ratios to estimate employment and wages. Employment in t^{th} month is estimated by the product of the ratio employment and *t-1* th month employment. Wages in the current quarter are estimated by the product of wages ratio and previous quarter wages, in formulas

$$(3)\qquad \widehat{empl}_t = empl_{t-1} \cdot RE_{h(t)} \qquad t = 4,5,6$$

$$\widehat{wage}_{current} = wage_{previous} \cdot RW \ .$$

All estimates are rounded to integer values using random rounding. Definition of the method of random rounding is given in section 6 of the paper..

3.5 Nearest Neighbor Method: An establishment with data is to be imputed is called *recipient*. An establishment from which data is used to estimate missing data of the *recipient* is called *donor*. In nearest neighbor method for every *recipient* we find one *donor*. The *donor* establishment is found in the following way: The establishments are arranged by the order of cell and random employment assigned to each establishment. Every establishment in the cell is then assigned either a value of 1 or 2 based using MOD-2 for a dummy variable counter plus 1 to every establishment in the dataset. Two nearest establishments are found from a cell of the recipient as possible donors, first donor is just above and other is just after the recipient establishment.. Depending upon the MOD value of the recipient , the donor is selected that matched the recipients MOD value. The donor's month to month employment ratios for the current quarter are computed, and quarter to quarter wages ratio is computed using the following

$$(4) \qquad re_t(do) = \frac{empl_t(do)}{empl_{t-1}(do)} \qquad t = 4, 5, 6$$

$$rw(do) = \frac{wages_{current}(do)}{wages_{previous}(do)}$$

then employments and wages of the recipient are estimated

$$(5) \qquad \widehat{empl}_t(re) = empl_{t-1}(re) \cdot re_t(do) \qquad t = 4, 5, 6$$

$$\widehat{wage}_{current}(re) = wage_{previous}(re) \cdot rw(do).$$

Notice obvious similarity between pairs of formulas (2), (3) and (4), (5). The only difference is that in ratio method we use the whole set of establishments from recipient's cell to compute ratio. In nearest neighbor method only one establishment, the "closest" one is used.

In a case of the last reported value of employment or wages being 0, both methods, ratio and nearest neighbor give estimates equal to 0. This outcome might be undesirable since many establishments with last reported value 0 in the previous quarter might turn into positive employment or wages in the current quarter. For this reason both methods, ratio and nearest neighbor, should be recommended for imputations on positive units only. Units which are not positive units, i.e. having 0 as the last reported value of employment, or 0 as the last reported wages should follow other method of imputation. Imputation methods for such units are not presented in this article.

4. Testing Nearest Neighbor and Ratio Methods

For every two consecutive quarters of data between Jan. 2000 and June 2007, we select a sample of reported units, as described in the outline of methodology, with cells defined by state / 2-digit NAICS / size class. In total we have 30 samples. The selected samples do not contain units with 0 employment in the third month of the previous quarter (since these units were recommended to

follow a different method of imputations). Imputations by nearest neighbor and ratio methods were used for estimating 90 months of employment and 30 quarters of wages.

We performed imputations by nearest neighbor and ratio methods using two types of cells structures:
n1 structure: cells defined by state, first digit NAICS, and size class
n2 structure: cells defined by state, first two digits NAICS, and size class

If n1 structure is used then the whole industry is divided by 9 sub-industries. The n2 structure gives 23 sub-industries. Obviously, n1 cells are larger than n2 cells, but n2 cells are more homogenous.

4.1 Error Measures: Imputed values of employment and wages are compared to actual values. The difference between imputed and actual is an error. We aggregate errors over states, nationwide by size class, and total nationwide. Express as percent we define relative error as a difference between the sum of imputed and actual values divided by the sum of actual

$$\text{Percent Relative Error} = 100 \times \frac{\sum imputed - \sum actual}{\sum actual} \tag{6}$$

where summations are performed on a sample over state, nationwide by size class, or nationwide.

4.2 Results: Table 2 contains statistic computed from percent relative errors from imputations by nearest neighbor (NN) and ratio method using two different cells structures n1, and n2. Percent relative errors are computed over all nationwide samples.

Table 2 Percent relative errors Statistics from employment and wages imputations, Jan. 2000 – Jun. 2007.

	Employment Per. Rel. Err.			Wages Per. Rel. Err.		
	Average	Median	St. Dev.	Average	Median	St. Dev.
n1_NN	-0.25	-0.23	0.33	5.17	6.35	5.27
n2_NN	-0.23	-0.25	0.33	5.73	6.53	2.78
n1_ratio	-0.11	-0.02	0.55	0.14	-0.02	1.68
n2_ratio	-0.19	-0.08	0.51	-0.01	0.01	0.93

Here n1_NN stands for nearest neighbor imputations with cell structure with first digit NAICS, n2_NN nearest neighbor imputations with first 2-digit NAICS, similarly, n1_ratio, n2_ratio.

5. Determining cut off values for influential observations

Looking at the data in Table 1, one might well ask how were these cutoffs points determined. Defining a reasonable set of cut off values for detection of influential observation is very important for the accuracy of the ratio method imputation. Different cut off values generally, produce varying accuracy of imputations for both employment and wage data. We provide a short outline of the research conducted for deriving criteria to deem an establishment as an influential observation. For finding cut off values we used only ratio method. The nearest neighbor method when applied with the same cut off values produced similar improvement in accuracy. The derivation consisted of about 400 simulations with random samples using different cut off criterion for influential observations. We tested all 4 quarters of data in the years 2001-2006. We looked for criteria that would give us a consistently low relative error of 1% or less at the national size class levels while limiting the number of influential observations to be less than 1.5% for all 7 years of data.

Initially, we assumed all cut-off values of 50. We began the simulation by: 1) selecting a random sample from the set of reported positive establishments that mimics the distribution of actual non-respondents; 2) computing the three employment ratios and the wage ratio for the current quarter of data at the State / 1-digit NAICS / Size Class cell level based on all non-sampled establishments excluding influential observations (i.e., ratios greater than 50); 3) applying these computed cell ratios to impute employment and wages for the random sample of establishments; 4) computing errors by subtracting the actual values from the imputed values for each establishment in the random sample; 5) summing individual errors to the State /1-digit NAICS / Size Class, and National / 1-digit NAICS / Size Class; 6) comparing the relative errors at the National / Size Class levels against the predefined criteria of 1%; and 7) comparing the number of influential observations against the predefined criteria of 1.5%.

The cut-off values are derived at the national/size class level. The results showed the relative errors on wages differ significantly from size class to size class. These cut off values for imputation in the first quarter of 2005 resulted in the percent relative errors detailed in the table below:

Table 3. Results of imputations with cut off value for influential observations equal to 50 for both employment and wages. Imputations performed into the first quarter of 2005.

	Percent Relative Error				
class size	month1	month2	month3	wages	%influential observations
(1) 1 - 4	5.00	5.29	4.22	6.24	0.08
(2) 5 - 9	2.00	2.14	1.51	-0.57	0.02
(3) 10 - 19	1.08	1.24	1.02	0.65	0.01
(4) 20 - 49	0.31	0.43	0.23	0.65	0.01
(5) 50 - 99	-0.06	0.06	0.01	0.11	0.01
(6) 100+	0.34	0.27	0.51	0.62	0.01
All Sample	1.34	1.44	1.21	1.32	0.05

The very high cut off value of 50 resulted in extremely small percentages of influential observations being removed - only 0.05% - and that is a very desirable outcome. However, we rejected this cut off value because of the following reasons:
1) Relative errors for employment for size classes 1, 2, 3 exceeded the pre-defined criteria and for wages for size class 1.
2) There was a significant bias in estimation. Almost all positive percent relative errors show overestimation.

These problems were persistent for all other quarters. Therefore, we rejected these cut off values for imputation. Using cut off values larger than 50 might even worsen the above listed problems. In particular, not removing any influential observations at all is likely to give worse results.

As a next example, we assume all cut off values to be very small, equal to 2, (i.e. every establishment which increased employment or wages 2 times is an influential observation). Imputing with all cut off values equal to 2 resulted in the percent relative errors as shown in the following table.

Table 4. Results of imputations with cut off value for influential observations equal to 2 for both employment and wages. Imputations performed into the first quarter of 2005.

	Percent Relative Error				
class size	month1	month2	month3	wages	% influential observations
(1) 1 - 4	-0.17	-1.38	-3.80	-3.83	12.11
(2) 5 - 9	1.06	0.69	-0.52	-4.98	3.56
(3) 10 - 19	0.53	0.35	-0.24	-3.13	2.63
(4) 20 - 49	-0.16	-0.23	-0.65	-3.64	1.94
(5) 50 - 99	-0.36	-0.44	-0.63	-4.03	1.68
(6) 100+	0.10	-0.07	0.03	-2.97	1.62
All Sample	0.13	-0.22	-0.89	-3.53	7.14

We rejected these cut off values for the following reasons:
1) Too many establishments were removed as influential observations, particularly in size class 1.
2) The percentage relative errors were very large for wages and showing a strong downward bias.

The described problems were persistent in all other quarters. We repeated these types of simulations about 400 times working with different quarters of data. Thus, between a few unacceptable extremes there existed a point at which the cut off values give both acceptable level of accuracy and an acceptable percentage of removed influential observations. We determined that point empirically and arrived at the values presented in the Table 1. These values most consistently provided acceptable levels of percent relative errors and percentage of observations classified as influential.

Table 5. Results of imputations with cut off value for influential observations as given in table 1 for both employment and wages.

	Percent Relative Error				
class size	month1	month2	month3	wages	% influential observations
(1) 1 - 4	0.09	0.12	0.07	-0.03	0.88
(2) 5 - 9	0.16	0.09	0.16	-1.41	0.59
(3) 10 - 19	-0.01	0.09	-0.05	-0.72	0.61
(4) 20 - 49	0.03	-0.02	-0.12	0.12	0.36
(5) 50 - 99	-0.06	0.13	0.07	-0.46	0.33
(6) 100+	0.36	0.10	0.00	0.28	0.45
All Size classes combined	0.14	0.09	0.01	-0.15	0.77

6. Random Rounding

Random Rounding: Let *x* be a real number that needs be rounded to an integer value by random rounding. Let *Int(x)* denote the largest integer smaller than *x*. Let *Ran(x)* denote the integer to which *x* is to be rounded by random rounding. We perform random rounding by randomly selecting a number *y* from uniform distribution on [0, 1], and using the formula:

$$Ran(x) = \begin{cases} int(x);\ y \geq x - int(x) \\ int(x) + 1;\ y < x - int(x) \end{cases}$$

7. Summary

Recommendations given by BLS, Statistical Methods Division: Nearest neighbor method was rejected due high errors for imputations of wages. Ratio method was recommended with n2 cell structure. Both n1 and n2 cell structure in ratio method perform almost equally for employment, but n2 is slightly better in wages. Similar comparisons of imputation methods were performed with statistics on nationwide size classes and statistics on states with similar results. The results for states are not included in this paper because of voluminous tables and may be requested from the authors.

References

Sandra West, Shail Butani, Micheal Witt, Craig Adkins, *Alternative Imputation Methods for Employment Data,* Sandra West (1990), Bureau of Labor Statistics, 2 Mass Avenue #4985NE, Washington, DC 20212

Jill M. Montaquila, Westat, Inc., and Chester H. Ponikowski, *An Evaluation of Alternative Imputation Methods* (1995), Chester Ponikowski, Bureau of Labor Statistics, Postal Square Building Suite 3160, 2 Massachusetts Ave., NE, Washington, DC 20212

Yves Thibaudeau, *Model Explicit Item Imputation for Demographic Categories*, Survey Methodology (2002), Vol. 28, No. 2, pp. 135-143, Statistics Canada