An Evaluation of BLS Noise Research for the Quarterly Census of Employment and Wages

Y. Michael Yang[1], Ali Mushtaq[2], Santanu Pramanik[3] Fritz Scheuren[4],
David Hiles[5], Michael Buso[6], Shail Butani[7]
[1234]NORC at the University of Chicago, 4350 East-West Highway, 8th Floor, Bethesda, MD 20814
[567]Bureau of Labor Statistics, Postal Square Building, 2 Massachusetts Avenue, NE
Washington, DC 20212

Abstract

The Quarterly Census of Employment and Wages (QCEW) program of the Bureau of Labor Statistics (BLS) publishes tabulations of monthly employment, quarterly wages, and number of establishments by industry and geography. In accordance with BLS policy, data provided to the Bureau in confidence should be used only for statistical purposes. In particular, the publication of data collected from BLS surveys should not lead to the identification of cooperating respondents. The BLS has been concerned about the current cell suppression method used with the QCEW. To address such concerns, BLS has conducted research about the random noise method as an alternative method. This paper provides an assessment of the BLS research to date. The paper begins with a review of the current cell suppression method in Section 2, focusing on the major disadvantages of the current method. Section 3 reviews the random noise model, its current application to QCEW at BLS, and the results to date. Section 4 provides an assessment of the properties of the BLS noise model under certain distributional assumptions of the noise factors. Finally, Section 5 provides some concluding remarks.

Key Words: Disclosure limitation, Cell suppression, Random noise method

1. Introduction

The Quarterly Census of Employment and Wages (QCEW) program of the Bureau of Labor Statistics (BLS) publishes tabulations of monthly employment levels, quarterly wages (total and average weekly), and number of establishments by industry and geography. In accordance with BLS policy, data provided to the Bureau in confidence should be used only for specified statistical purposes. In particular, the publication of data collected from BLS surveys should not lead to the identification of cooperating respondents. The BLS has been concerned for some time about the current cell suppression method used with the QCEW program. To address such concerns, BLS has conducted research about the random noise method as an alternative method for the QCEW. This report provides an assessment of the BLS research to date. The report begins with a review of the current cell suppression method in Section 2, focusing on the major disadvantages of the current method. Section 3 reviews the random noise model, its current application to QCEW at BLS, and the results to date. Section 4 provides an assessment of the BLS noise model under certain distributional assumptions of the noise factors, including an investigation of the effect of stacking two multiplicative noise factors. Finally, Section 5 provides some concluding remarks.

2. The Cell Suppression Method

2.1 The Current Cell Suppression Method

QCEW releases tabular data on a quarterly basis, where tabulation is done by industry and geography. Industry detail is at the 6-digit North American Industry Classification System (NAICS) level, which involves over 1,200 detailed industries. Meanwhile, higher levels of industry aggregation are also used for tabulations. Overall, there are a total of nearly 2,400 industries at various levels of aggregation. Geographic detail is at the county, Metropolitan Statistical Area (MSA), state, and national levels, for a total of nearly 4,000 areas. In addition to industry and geography, ownership and size of the establishment are also used in tabulations. In terms of data types, QCEW releases data on establishment count, total wages, taxable wages, contributions, and average weekly wages. Therefore, for a particular data release, a QCEW cell is defined by a particular data type for a particular combination of industry, area, ownership, and size. Overall, there are approximately 3.6 million non-empty QCEW cells for each data type.

Current QCEW data release is based on the cell suppression method to reduce disclosure risk. The cell suppression method involves two phases: primary suppression and secondary suppression. The first phase assesses the sensitivity of the data in each cell, based on the number of establishments in the cell and the dominance of the larger establishments in the cell. This is done using a "p-percent" rule, which effectively tests whether there are only 1 or 2 dominant establishments for a particular sensitive variable (e.g., total wages) within a cell. Sensitive cells that present significant disclosure risks are suppressed.

2.2 Disadvantages of the Cell Suppression Method

The most obvious disadvantage of the cell suppression method is that it has resulted in a large amount of data being suppressed, compromising the quality and utility of the QCEW data products. In particular, this method suppresses much information that is not at risk for disclosure. Any cell that is used as a complementary suppression represents data that could have been published if there were other ways of protecting the sensitive cells. QCEW data are in great demand, not only for the current data products, but also in greater detail. For example, tabulations for sub-county areas will be very useful for policy studies involving legislative districts, cities, central business districts, and so on. However, data publication for more detailed geographic areas will be subjected to even higher suppression rates under the current cell suppression method.

The cell suppression method is also difficult to implement in practice, especially for a program like QCEW that releases a large amount of data on a regular basis. While determining primary suppression cells are relatively straightforward, the process of choosing complementary suppressions is very complicated and time consuming. Complementary suppressions must be performed separately for each data product, which means that the analysts must keep track of the suppression patterns among all tabulations. The current QCEW complementary suppression process adds two to three weeks to the data publication process.

Finally, the cell suppression method may be of questionable adequacy by state-of-the-art standards. The techniques used by QCEW emulate the techniques that a user would do manually for a small table. They take into account many dimensions and levels of aggregation and would be impossible to actually do manually. They are, however, potentially vulnerable to higher levels of mathematical attack. With sufficient computing power, a data attacker could feed all of the QCEW data into a program and solve for at least some of the sensitive cells. Therefore, even with extensive suppressions, the current method may not provide sufficient protection from sophisticated attacks.

3. The Random Noise Method

In recent years, the random noise method has been gaining wider use in statistical agencies to protect respondent data from unintended disclosure. The QCEW program has been conducting research on the use of the random noise method as an alternative to the cell suppression method.

3.1 The Original Noise Method

The original random noise method was developed in the late 1990s by Tim **E**vans, Laura **Z**ayatz, and John **S**lanta (Evans, Zayatz, and Slanta, 1998). The so-called **EZS** noise method takes a *micro* approach to disclosure limitation: a multiplier, or noise factor, is applied at the unit level rather than at the cell level. Under this method, a noise factor is applied to each unit prior to any tabulation, which guarantees that different tabulations, from the lowest to the highest level, are consistent. Applying the **EZS** method to QCEW data, noise would be added to each establishment's employment and wage data. Each establishment would be assigned a different noise factor and each of its reported values is multiplied by this factor. For example, noise factors of 1.1 or 0.9 would perturb the true value by 10% in either direction. Multiplying, as opposed to adding, creates a perturbation proportional to the true value, which is one innovative aspect of the **EZS** approach. The size of the perturbation depends on the magnitude of the original values (Cox and Dandekar, 2003; Strudler, Oh, and Scheuren, 1986). Ten percent is used throughout the **EZS** literature.

The motivation for the **EZS** method is that cells with a large number of establishments have a low risk of disclosure. If the number of establishments is large enough, the reported cell aggregate should be close to the actual value because the noise should roughly cancel out. For sensitive cells, those with a small number of establishments or those dominated by a single establishment, the noise will not cancel out and hence the noisy aggregate will represent relatively larger perturbation. This will prevent users from being able to recover an individual respondent's true value from the published value (EZS 1998). In theory, the effect of the added noise is only substantial in cells where confidentiality is at risk.

Under the standard **EZS** method, the cell aggregate is unbiased since the expected value of the multiplicative noise factor is 1. The variance of a cell aggregate may be studied by simulating several different factor assignments. Once the noise has been added, cells may be raked to the appropriate totals and tabulations can be carried out. Even with raking, the simplicity of the **EZS** method is a significant advantage, especially compared to complementary cell suppression. The cells have been protected and no suppressions are necessary.

3.2 The QCEW Noise Method

BLS has performed research on noise methodologies based on EZS, with some modifications. The EZS method by itself does not address the peculiarities in the QCEW data. Some enhancements were made to apply the idea behind the EZS method to the QCEW data. In this section, after introducing some notations, we describe the BLS noise model step by step with emphasis on the modifications of the EZS method.

For a particular cell (cell is defined by industry and region, to avoid complication we suppress the cell identifier), let's define, E_{ij} = Employment level of the *j*th establishment belonging to the *i*th company, TW_{ij} =Total wage of the *j*th establishment belonging to the *i*th company, $R_{ij} = TW_{ij} / E_{ij}$, Average wage of the *j*th establishment belonging to the *i*th company, $j = 1,..., N_i$, N_i being the number of establishments belonging to the *i*th company, $i = 1,..., C$, C being the total number of companies in that particular cell. At the aggregate level (i.e., cell level), total employment is $E = \sum_{i=1}^{C} \sum_{j=1}^{N_i} E_{ij}$, total wage is $TW = \sum_{i=1}^{C} \sum_{j=1}^{N_i} TW_{ij}$, and hence, average wage is $R = TW/E$. For the QCEW program, BLS publishes the employment statistics (E) on a monthly basis and the wage statistics on a quarterly basis. To publish the average wage on a quarterly basis, in the denominator of R , BLS uses the average monthly employment level (i.e., average of the employment totals corresponding to three months of a quarter).

Step1: Apply Noise to the Employment Variable

To add noise to the employment variable, BLS assumes the model $E_{ij}^{*} = E_{ij} \varepsilon_{1i} \varepsilon_{2ij} + \varepsilon_{3ij}$, where ε_{1i} is the random multiplicative noise at the company level (the same noise is used for all the establishments belonging to the company), ε_{2ij} is the random multiplicative noise at the establishment level, and ε_{3ij} is the random additive noise at the establishment level.

The original EZS method does not recommend any additive noise. An additive noise is considered for the QCEW employment level data because the values are reported as discrete whole numbers, and many of the employment values are small or even zero. Since no multiplicative noise factor can protect a value of zero, and only an extreme range of noise factors can protect values of one and two, BLS concluded that something other than or in addition to multiplicative noise was needed to protect the small values from disclosure.

Another enhancement BLS made to the EZS method is to incorporate two multiplicative noise factors instead of one. One noise factor is at the company level and the other at the establishment level. The application of two noise factors serves two purposes. First, it helps to protect establishments within the same company. If only the company level noise factor were considered, different establishments in the same company would have the same base noise factor and one establishment could accurately estimate the true values for the other establishments in the same company, particularly for small cells. While one would generally expect that establishments in the same company might not need to protect their information from other establishments of that company, this might not always be the case. This also has the advantage of protecting against outside attack. If an attacker were to obtain from other sources enough information about a single establishment that he could closely bracket the effective noise factor for that establishment, he would not be able to transfer that knowledge to tightly approximate the data for the other establishments of that company. Second, it helps to protect the company level data. If only the establishment level noise factor is used, the noise will tend to cancel out and the perturbations only at the establishment level will result in almost no perturbation at the company level, particularly for large companies consisting of many establishments. After applying this model, BLS obtains the fuzzy employment total for the cell as $E^* = \sum_{i=1}^{C}\sum_{j=1}^{N_i} E_{ij}^*$.

Step2: Apply Noise to the Average Wage Variable

To add noise to the average wage variable, BLS assumes the model $R_{ij}^* = R_{ij} e_{1i} \varepsilon_{2ij}$, where e_{1i} is the random multiplicative noise at the company level, and ε_{2ij} is the random multiplicative noise at the establishment level. Again, the same noise factor is applied to all the establishments belonging to the company. Note that the company level multiplicative noise factor (e_{1i}) is different than what is considered for the employment variable (ε_{1i}) but the same establishment level multiplicative noise (ε_{2ij}) is considered for both variables. No additive noise is applied for the average wage variable.

Step3: Apply Noise to the Total Wage Variable

To add noise to the total wage variable, BLS proceeds as follows:

$$TW_{ij}^* = R_{ij}^* \times E_{ij}^* = \left(R_{ij} e_{1i} \varepsilon_{2ij}\right) \times \left(E_{ij} \varepsilon_{1i} \varepsilon_{2ij} + \varepsilon_{3ij}\right)$$

$$= R_{ij} E_{ij} \varepsilon_{1i} e_{1i} \varepsilon_{2ij}^2 + R_{ij} e_{1i} \varepsilon_{2ij} \varepsilon_{3ij}$$

$$= TW_{ij} \varepsilon_{1i} e_{1i} \varepsilon_{2ij}^2 + R_{ij} e_{1i} \varepsilon_{2ij} \varepsilon_{3ij}$$

After applying the above model, the fuzzy total wage for a particular cell is obtained as

$$TW^* = \sum_{i=1}^{C}\sum_{j=1}^{N_i} TW_{ij}^* .$$

Step 4: Calculate Fuzzy Average Wage

Once the fuzzy employment total and fuzzy wage total are obtained, BLS calculates the fuzzy average wage as $R^* = TW^* / E^*$.

These four steps summarize the BLS base noise model for the QCEW data. The base model is further complicated by implementation details, which are still being explored by the NORC team. Not all data is being treated by the noise model. Wage and employment statistics of the Federal government, for example, aren’t protected because it is public information. Also, some rounding occurs in the employment data to yield integer employment values. Any effect caused by rounding has to be examined to ensure no unintended bias results. Another modification that BLS intends to incorporate into the base noise model is to retain noise factors for establishments (and companies) over time. A small percentage of establishments will take on new random noise factors on a rolling basis. But to preserve time series data, noise factors will remain attached to establishment for some number of quarterly releases. While this idea has not been implemented in BLS research to date, the desire is to allow real time series trends within aggregated cells to still be visible in the noise-treated data.

3.3 Preliminary Results

Although BLS researchers have considered a few ingenious modifications to the original EZS method, they are somewhat disappointed by the preliminary results of their noise research. They observed three stumbling blocks in their work which prevented them from going further.

The first is that the noise model shifts high-level aggregates from the true values more than desired. The expectation is that, at high levels, noise would more or less cancel out so that the noise-treated estimates would be close to the original values for large cells. BLS observed that the difference between the noise-treated and original total is larger than desired. However, such difference is to be expected under the random noise method without post-noise treatment adjustments.

Another problem encountered during BLS research is negative aggregates caused by the additive noise factor. There are establishments with very low, even zero, reported employment. With an additive noise centered on zero, it is possible to see negative employment in the treated data, and hence aggregates, especially with low-level aggregates. This is a violation of the data constraints imposed on the QCEW data. In their current implementation, BLS allowed the negative values in their research at the micro level, so as not to cause bias due to rounding up all negative values. However, at the cell level BLS plans to truncate the negative aggregate value at 0. This procedure, however, may not preserve the additive property that exists between lower level aggregates (6-digit industry code by county) and higher level aggregates (6-digit industry code by state).

A third problem with the current implementation is bias. BLS researchers observed a systematic positive bias in the fuzzy totals after applying the noise model. The noise model, while complicated for some models, appears symmetric. But graphs of a few aggregates at high levels clearly show more positively affected cells than negatively affected ones. NORC investigated the source of this bias both theoretically and empirically, and the results of our investigation are discussed in Section 4.

4. Assessment of the BLS Noise Model

4.1 Statistical Properties of the BLS Noise Model

In this section, we investigate the bias and variance of the noise-treated totals under the BLS noise model, given certain distributional assumptions regarding the error terms. Although BLS did not consider exactly the same distributional assumptions for the error terms, they considered similar type of symmetric random noise terms for the QCEW data and hence the results will be similar. This is an effort to investigate the bias that BLS researchers noticed in their implementation of the noise model.

Bias and Variance of Fuzzy Employment Total E^*

Let's assume mixture normal distribution for the multiplicative noise factors and normal distribution for the additive noise factor. That is, we assume

$$\varepsilon_{1i} \sim \alpha_1 N\left(\mu_{11}, \sigma_1^2\right) + \left(1-\alpha_1\right) N\left(\mu_{12}, \sigma_1^2\right)$$

$$\varepsilon_{2ij} \sim \alpha_2 N\left(\mu_{21},\sigma_2^2\right)+\left(1-\alpha_2\right)N\left(\mu_{22},\sigma_2^2\right)$$

$$\varepsilon_{3ij} \sim N\left(\mu,s^2\right),$$

where $N\left(\mu,\sigma^2\right)$ denotes normal distribution with mean μ and variance σ^2 (standard deviation σ), α_1 and α_2 are the mixing parameters. We also assume that the error terms are independent of each other. For the multiplicative noise factors, a sensible assumption is to consider the mean values of normal distributions to be closer to 1 and in opposite directions (e.g., $\mu_{11}=1.1$ and $\mu_{12}=0.9$) and the mixing parameter (α_1) to be 0.5 in order to be able to perturb 50% of the data in one direction and 50% in the other direction. The variance parameters should be small (e.g., $\sigma_1=0.02$) to avoid extreme perturbation of the original values. For the additive noise factors, it makes sense to add noise symmetrically about 0 (i.e., $\mu=0$) to avoid bias in either direction. Again the value of the dispersion parameter should be small. For example, the value of $s=2$ will add noise lying between -4 and +4 to the true employment values in 95% of the cases.

Under the above distributional assumptions, we obtain the expectation of E^* as

$$E\left(E^*\right)=\sum_{i=1}^{C}\sum_{j=1}^{N_i}E_{ij}\left[\left\{\alpha_1\mu_{11}+\left(1-\alpha_1\right)\mu_{12}\right\}\left\{\alpha_2\mu_{21}+\left(1-\alpha_2\right)\mu_{22}\right\}+\mu\right]$$

$$=\sum_{i=1}^{C}\sum_{j=1}^{N_i}E_{ij}=E$$

The last step is true when $\alpha_1=\alpha_2=0.5$, $\mu_{11}+\mu_{12}=2$, $\mu_{21}+\mu_{22}=2$, and $\mu=0$. All these are reasonable assumptions, as discussed above. So, we conclude that the fuzzy employment total is an unbiased estimator of the true employment total.

The variance of the fuzzy employment total E^* is given by

$$Var\left(E^*\right)=\left[Var\left(\varepsilon_{1i}\right)Var\left(\varepsilon_{2ij}\right)+Var\left(\varepsilon_{1i}\right)+Var\left(\varepsilon_{2ij}\right)\right]\sum_{i=1}^{C}\sum_{j=1}^{N_i}E_{ij}^2+s^2,$$

where $Var\left(\varepsilon_{1i}\right)=\sigma_1^2+\alpha_1\left(1-\alpha_1\right)\left(\mu_{11}-\mu_{12}\right)^2$ and $Var\left(\varepsilon_{2ij}\right)=\sigma_2^2+\alpha_2\left(1-\alpha_2\right)\left(\mu_{21}-\mu_{22}\right)^2$.

The mathematical details are not presented here, but are available upon request. From the above expression, we observe that the greater the difference between the two mean parameters of the mixture normal distribution, the greater the variance of E^*. We have noted earlier that as long as the sum of the two mean parameters is 2 (with mixing parameter being 0.5), E^* is an unbiased estimator of E. To reduce the variance, we have to be careful about the choice of the mean parameters. For example, the choice of $\mu_{11}=1.1$ and $\mu_{12}=0.9$ is preferable to $\mu_{11}=1.2$ and $\mu_{12}=0.8$ if variance reduction is an issue (although both the choices lead to unbiased estimator, since $\mu_{11}+\mu_{12}=2$).

Another interesting point is worth mentioning. The expression $\alpha\left(1-\alpha\right)$ in the variance terms above is maximum when $\alpha=0.5$, and ironically that's the value we recommend as mixing parameter in order to get an unbiased estimator (fuzzy totals). We think that's a reasonable choice as our main focus is to find an unbiased estimator in order to reflect the true picture of the employment and wages. Also from the disclosure control perspective, it may sometimes be necessary to add variability to the fuzzy totals. So our choice of mixing parameter, corresponding to the mixture normal distribution, serves both the purposes. Even with the choice of $\alpha=0.5$, variance reduction is possible (if that's a criteria) by choosing the normal distribution means accordingly, as discussed above.

Bias of Fuzzy Total Wages TW^*

Under the same mixture normal distributional assumption for the multiplicative noise factors and normal distribution for the additive noise factor, the expectation of TW^* is given by

$$E\left(TW^*\right)=\sum_{i=1}^{C}\sum_{j=1}^{N_i}E\left(TW_{ij}^*\right)=\sum_{i=1}^{C}\sum_{j=1}^{N_i}E\left\{TW_{ij}\varepsilon_{1i}e_{1i}\varepsilon_{2ij}^2+R_{ij}e_{1i}\varepsilon_{2ij}\varepsilon_{3ij}\right\}$$

$$=TW\left[Var\left(\varepsilon_{2ij}\right)+1\right]+\sum_{i=1}^{C}\sum_{j=1}^{N_i}R_{ij}E\left(\varepsilon_{3ij}\right)$$

$$=TW\left[Var\left(\varepsilon_{2ij}\right)+1\right]$$

The last two steps follow from the assumption of the mixing parameter being 0.5, the sum of the two normal means being 2, the mean of the additive normal noise being 0, and the independence of the noise factors. Therefore, TW^* is not an unbiased estimator of TW. The positive bias arises because the same establishment level multiplicative noise factor (ε_{2ij}) is used for both employment (E) and average wage (R) variables.

The bias is zero only when $Var\left(\varepsilon_{2ij}\right)$ is zero, but that's not realistic. In other words, the current BLS procedure would always incur a positive bias while estimating the total wages. The bias would disappear only if the establishment level multiplicative noise factors for the two variables are independent.

Bias of Fuzzy Average Wage R^*

The noise-treated average wage R^* is a nonlinear estimator, being the ratio of two estimators TW^* and E^*. So to calculate the expectation of R^*, we need to apply Taylor series linearization technique. After some simplification, we find (details are available upon request),

$$E\left(R^*\right)\approx R\left[Var\left(\varepsilon_{2ij}\right)+1\right].$$

This approximation follows from the fact that E^* is an unbiased estimator of E but TW^* is not an unbiased estimator of TW, with the amount of bias being $TW\left[Var\left(\varepsilon_{2ij}\right)\right]$. Therefore, the current BLS procedure would always incur a positive bias in estimating average wage as well.

4.2 A Simulation Study

To bolster the theoretical results discussed above, we performed a simulation study to evaluate the performance of the BLS noise model based on a small dataset containing variables similar to that of the QCEW data. Our dataset contains 75 establishments distributed over 25 companies with company size varying from 1 to 15. We divided the 75 establishments into 15 cells with cell size ranging from 2 to 15. We then generated the employment level variable (E) using a gamma distribution with shape and scale parameters chosen to allow for a large variability. Specifically, we chose shape=0.5 and scale=75, which essentially generates values from a gamma distribution with mean 37.5 and variance 2812.5. The total wage (TW) variable is also generated from a gamma distribution with mean and variance depending on E values. In general, the greater the value of E, the greater the value of TW. Among the 15 cells, 9 are identified as sensitive cells. Out of the 9 sensitive cells, 6 are sensitive because these have only 2 establishments per cell. The remaining 3 are sensitive based on the p-percent rule with respect to the total wage variable. We consider the 10% rule, which means that if the combined total wage of a cell, after excluding the first and the second largest establishments (in terms of the employment size) is less than 10% of the largest establishments' total wage, it should be treated as sensitive. Once the data is created, we consider it to be fixed and apply the noise model (step1-4) 10,000 times to generate simulated fuzzy values 10,000 times. As discussed earlier, we assumed a mixture normal distribution for the multiplicative noise factors and normal distribution for the additive noise factor.

Simulation Results

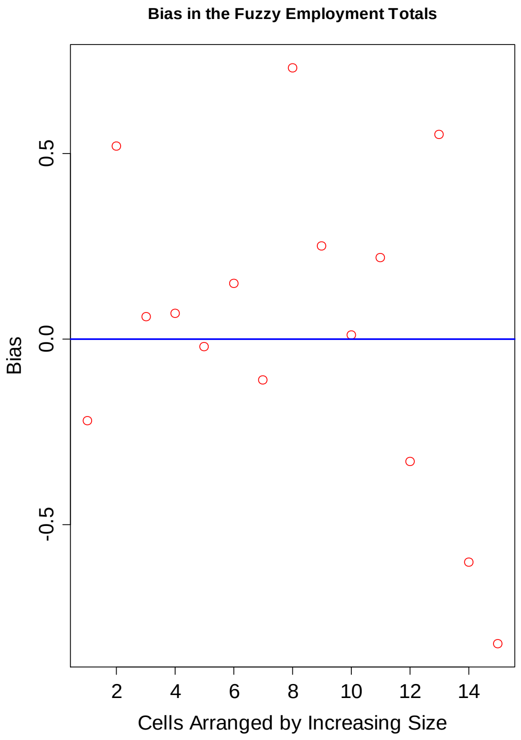
To evaluate the BLS noise model (along with our distributional assumptions regarding the noise factors), we define the following two summary statistics:

$$Bias\left(T^*\right) = E\left(T^*\right) - T$$

$$Ratio = E\left(T^*\right)/T,$$

where T^* is any fuzzy total and T is the corresponding true value. In our situation T^* can be E^*, TW^*, R^*. The expectation is taken over the simulated draws.

First we plot the bias and ratio for the employment totals. We arrange the cells by the size of the number of establishments in it. In other words, in the plot the first cell (in x-axis) is the smallest (of size 2) and the last cell (15th) is the largest (of size 15). From the bias plot we can see that the values are close to the zero line for all the cells and randomly distributed on the both sides of the zero line. Similar conclusion can be drawn from the ratio plot, and this time it is along the 1-line, as it should be. The plots support our theoretical finding that fuzzy employment total is an unbiased estimator of the true total.


Bias in the Fuzzy Employment Totals

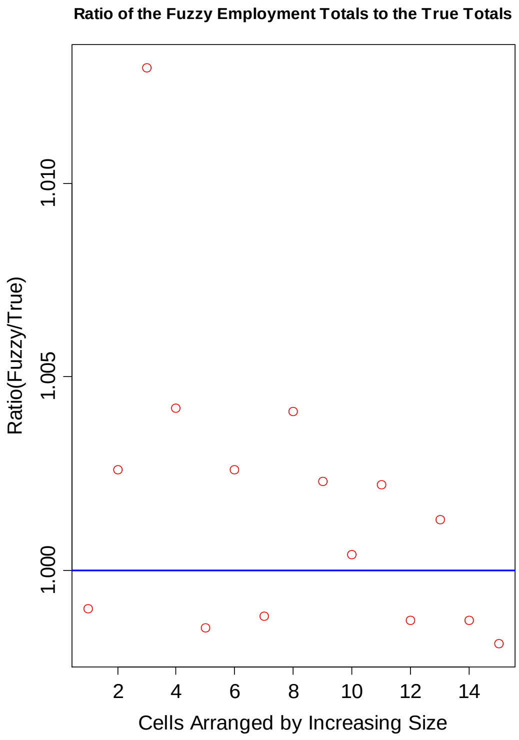

Ratio of the Fuzzy Employment Totals to the True Totals

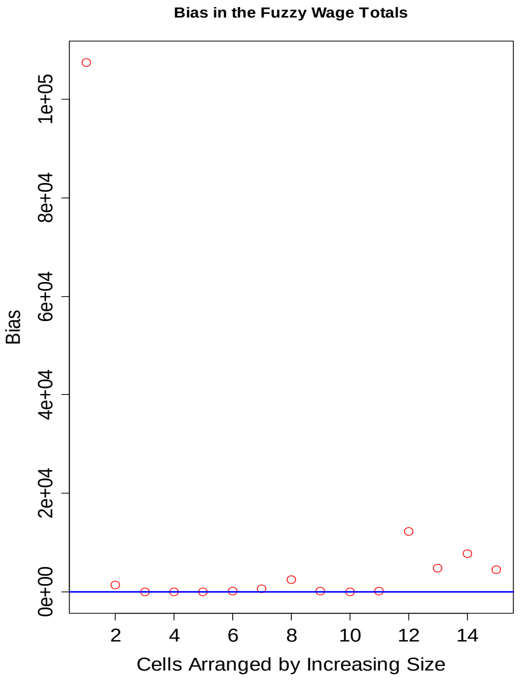
Table1. Fuzzy Employment Totals Along with 95% Confidence Intervals for 15 Cells

Cell	No.Estb	Sensitive	True Total	Fuzzy Total[1]	LCL*	UCL	Bias	Ratio

Cell 2	2	Yes	205	205	141	279	-0.22	0.9990
Cell 3	2	Yes	198	199	141	263	0.52	1.0026
Cell 5	2	Yes	5	5	0	11	0.06	1.0130
Cell 8	2	Yes	16	16	8	25	0.07	1.0042
Cell 11	2	Yes	15	15	7	24	-0.02	0.9985
Cell 14	2	Yes	59	59	39	81	0.15	1.0026
Cell 9	3	Yes	91	91	62	123	-0.11	0.9988
Cell 7	4	No	177	178	125	236	0.73	1.0041
Cell 15	4	Yes	112	112	77	151	0.25	1.0023
Cell 1	5	No	23	23	12	34	0.01	1.0004
Cell 13	5	No	97	97	67	130	0.22	1.0022
Cell 4	6	Yes	249	249	189	316	-0.33	0.9987
Cell 12	9	No	435	436	326	555	0.55	1.0013
Cell 10	12	No	445	444	338	562	-0.6	0.9987
Cell 6	15	No	431	430	324	547	-0.82	0.9981

[1] Fuzzy Totals are the mean of the 10,000 simulated fuzzy totals.
*Lower Confidence Limit. LCL and UCL are the 2.5% and 97.5% quantile.

Next, we plot the bias and ratio for the total wage variable. We see that almost all the bias values are plotted above the zero line and almost all the ratio values are above the 1-line. This is consistent with our theoretical result that suggests an upward bias in the estimates of total wage.



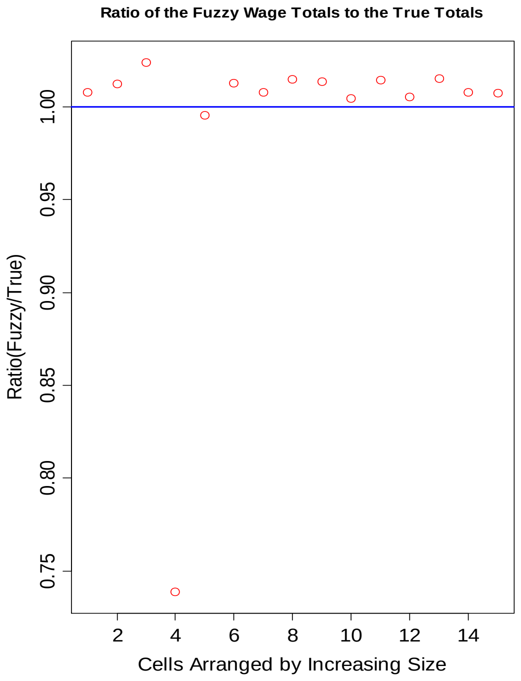


Table2. Fuzzy Total Wages Along with 95% Confidence Intervals for 15 Cells

Cell	No.Estb	Sensitive	True Total	Fuzzy Total[1]	LCL*	UCL	Bias	Ratio

Cell 2	2	Yes	13499693	13607172	7508626	22334648	107479	1.0080
Cell 3	2	Yes	115968	117412	68510	184531	1444	1.0125
Cell 5	2	Yes	197	202	0	444	5	1.0240
Cell 8	2	Yes	24	18	0	54	-6	0.7388
Cell 11	2	Yes	1361	1355	625	2342	-6	0.9954
Cell 14	2	Yes	8934	9047	4895	14861	113	1.0127
Cell 9	3	Yes	79567	80192	45364	128398	625	1.0079
Cell 7	4	No	164767	167233	95466	265013	2466	1.0150
Cell 15	4	Yes	5210	5280	3003	8358	70	1.0135
Cell 1	5	No	2563	2575	1165	4321	12	1.0045
Cell 13	5	No	14222	14428	7566	23810	206	1.0145
Cell 4	6	Yes	2223472	2235665	1255797	3651851	12193	1.0055
Cell 12	9	No	314404	319240	192914	490878	4836	1.0154
Cell 10	12	No	974343	982138	645270	1421167	7794	1.0080
Cell 6	15	No	611458	615997	350349	990195	4538	1.0074

[1] Fuzzy Totals are the mean of the 10,000 simulated fuzzy totals.
*Lower Confidence Limit. LCL and UCL are the 2.5% and 97.5% quantiles.

On the plot for the average wage variable, we notice a similar upward bias although it's not as dramatic as it is for total wages.

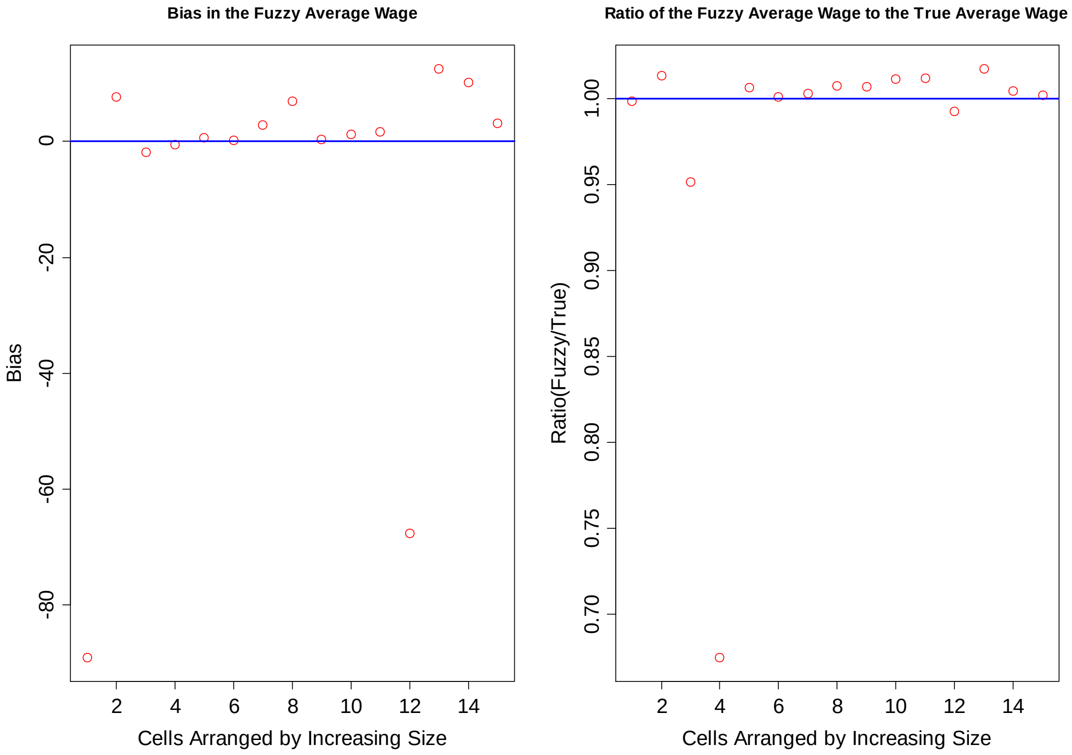


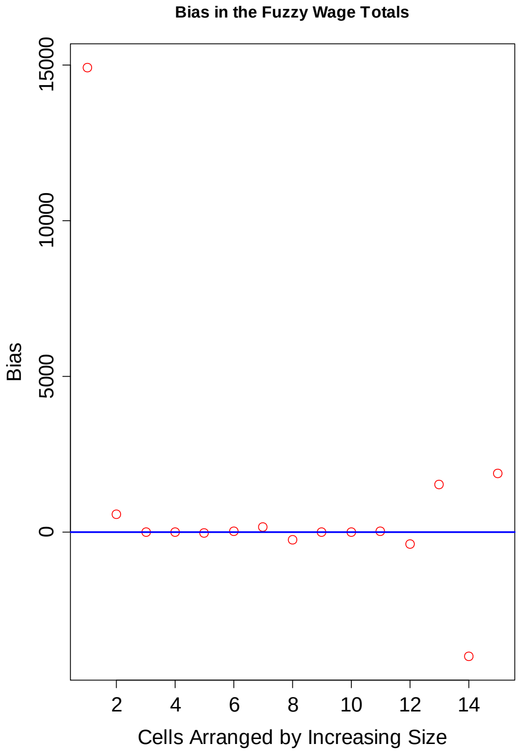
Table3. Fuzzy Average Wage Along with 95% Confidence Intervals for 15 Cells

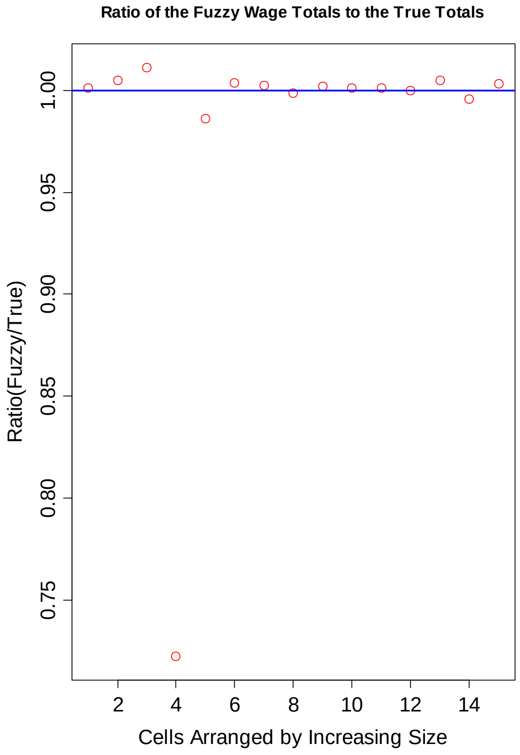
CellID	No.Estb	Sensitive	True	Fuzzy	LCL	UCL	Bias	Ratio

Cell 2	2	Yes	65852.2	65763.1	51011.4	82479.4	-89.06	0.9986
Cell 3	2	Yes	585.7	593.5	372.4	910.8	7.76	1.0132
Cell 5	2	Yes	39.4	37.5	0	51.9	-1.91	0.9516
Cell 8	2	Yes	1.5	1	0	2.8	-0.5	0.6746
Cell 11	2	Yes	90.8	91.3	62.8	129.8	0.58	1.0063
Cell 14	2	Yes	151.4	151.6	114.1	194.6	0.17	1.0011
Cell 9	3	Yes	874.4	877.2	678	1097.3	2.83	1.0032
Cell 7	4	No	930.9	937.9	691.9	1225.4	7	1.0075
Cell 15	4	Yes	46.5	46.8	36.4	58.3	0.33	1.0070
Cell 1	5	No	111.4	112.7	68.1	170.9	1.26	1.0113
Cell 13	5	No	146.6	148.3	98.4	211.5	1.71	1.0117
Cell 4	6	Yes	8929.6	8862	5960.6	12381.2	-67.6	0.9924
Cell 12	9	No	722.8	735.3	461	1116.6	12.56	1.0174
Cell 10	12	No	2189.5	2199.7	1740	2737.8	10.14	1.0046
Cell 6	15	No	1418.7	1421.9	1001.8	1923.6	3.17	1.0022

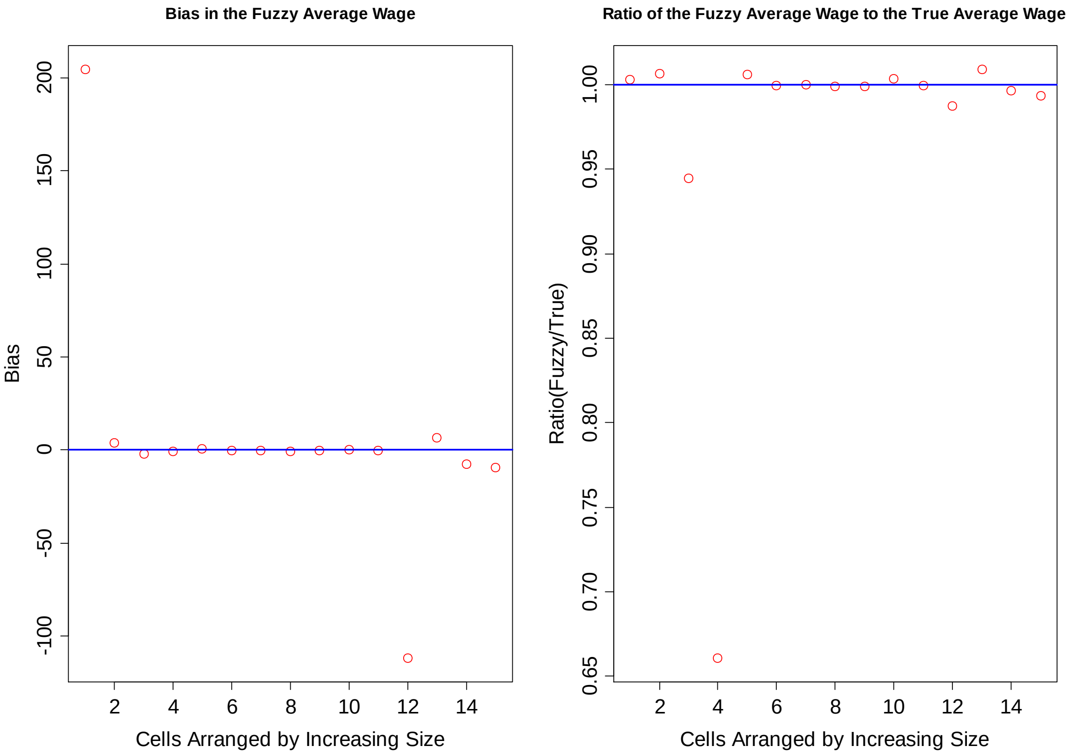
An Alternative to Remove Bias in Total Wage and Average Wage Estimates

To remove the upward bias in the noise-treated total wages and average wages, we could use two different multiplicative noise factors at the establishment level for employment (E) and average wage (R). After applying two independent multiplicative noise factors for two variables in our simulation, we obtained the following results.


Bias in the Fuzzy Wage Totals


Ratio of the Fuzzy Wage Totals to the True Totals

These plots show that the fuzzy totals are unbiased when the establishment level noise factors are different for employment level and average wage variable.



5. Concluding Remarks

We agree that the original EZS noise model cannot be used directly to protect the QCEW data without significant modifications. We commend BLS for its creative application of the EZS model to the QCEW data. Instead of a single multiplicative noise factor, BLS applied three noise factors: a company level multiplicative noise factor, an establishment level multiplicative noise factor, and an additive noise factor at the establishment level. These factors are designed to protect company level data, establishments within the same company, and establishments with very small number of employees. To avoid disclosure because of the mathematical relationship between the variables, BLS applied different noise factors for these variables. To preserve the validity of time series analysis, BLS proposed to use the same factor over time. In addition, BLS has considered further enhancements, such as considering alternative distributions for the noise factors, alternating the direction of the noise factors, and using raking procedures to control for over-adjustment in cell aggregates.

The inclusion of an additive noise factor is a major modification researched by BLS. The purpose of this noise factor is to protect small establishments that cannot be fully protected by a multiplicative noise. The application of two multiplicative noise factors is another major modification proposed by BLS. The purpose of these factors is to protect company level data as well as establishments within the same company. Because the QCEW is so comprehensive and collected at such regular intervals, it is important to consider how the base noise factors will affect the month-to-month and quarter-to-quarter relationships. Toward this goal, BLS proposed to use the same factors over time while updating the noise factor for 5% of the establishments. However, if the establishment base noise factor does not change from month-to-month or quarter-to-quarter, it might be possible for a user to correctly calculate the ratio between month-to-month or quarter-to-quarter values for a given establishment. It is important to avoid disclosing this information and it can be protected by adjusting the base noise factor each month, although not by too much.

In conclusion, we are very encouraged by the BLS research to date which we believe is fruitful and on the right track. As we have seen, the positive bias can be removed by using independent noise factors. Moving forward, we will work together to develop a nondisclosure model based on the random noise method. We will also be working on a raking model to deal with the issue of under perturbation and over perturbation. Most importantly, we will carry out extensive research to develop a nondisclosure model for cells with one or two employers.

References

Evans, B. T., Zayatz, L., and Slanta, J. (1998), "Using Noise for Disclosure Limitation for Establishment Tabular Data," *Journal of Official Statistics*, Vol. 14, No. 4, pp. 537-552.

Cox, L.H. and Dandekar, R. A. (2003), "A New Disclosure Limitation Method for Tabular Data that Preserves Data Accuracy and Ease of Use." Proceedings of the 2002 FCSM Statistical Policy Seminar, Federal Committee on Statistical Methodology.

Strudler, M., Oh, H. and Scheuren, F. (1986), "Protection of Taxpayer Confidentiality with Respect to the Tax Model." *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 375-381.