Using Data Mining to Explore Seasonal Differences Between the U.S. Current Employment Statistics Survey and the Quarterly Census of Employment and Wages October 2009

G. Erkens BLS
G. Erkens, Bureau of Labor Statistics, Office of Employment and Unemployment Statistics
Suite 4985, Postal Square Building, 2 Massachusetts Ave., N.E., Washington, DC 20212

Abstract

The Current Employment Statistics survey is a major economic indicator published by the Bureau of Labor Statistics to provide a timely measure of total payroll employment in the Nation. The Quarterly Census of Employment and Wages is a census of total employment that is available about 9 months after the CES estimate. While these two BLS programs measure the same information, they don't show the same seasonal patterns.

BLS recently conducted a response analysis survey (RAS) to ascertain why the patterns from the two programs differ, and this research employed classification tree algorithms to explore the data, using trees to select variables that exhibit a large influence on the seasonal difference. The results were more informative than previous searches through multiple sets of tabs, and several interesting and unconsidered relationships emerged.

Keywords: Response Analysis Survey, Classification and Regression Trees, Conditional Inference Trees.

1. Introduction

There are two BLS programs that measure total number of jobs at U.S. establishments: the Current Employment Statistics (CES) and the Quarterly Census of Employment and Wages (QCEW). The first is a sample of business establishments, while the latter is a census of business establishments.

Since each program offers a measure of total, national, private employment an observer might expect that the two programs offer similar employment totals and similar seasonal patterns. While the not seasonally adjusted employment totals are comparable, their seasonal patterns are sometimes distinctly different. Particular months offer different pictures of what is happening in the economy, and these pictures differ fairly consistently. Previous researchers have explored the seasonal differences between the two programs using a Response Analysis Survey (RAS), and while some results provided useful insights (Werking et al. 1995), a final picture of the underlying cause remained elusive.

The Bureau of Labor Statistics recently conducted another RAS to reinvestigate these seasonal differences. To analyze responses to the current RAS I employed a statistical

learning method known as Classification and Regression Trees (also known as CART) or recursive partitioning. I used these methods to explore the relationships and influences amongst the different responses. This paper consists of six sections. Section 2 briefly describes the two programs and the CES estimation methodology. Section 3 describes the seasonal differences between the two surveys. Section 4 briefly describes the previous and current RAS. Section 5 discusses the different statistical learning methods employed for this research. Section 6 discusses how these methods were applied and the results of this analysis. Section 7 is a summary that discusses future possibilities for research.

2. Employment Programs at BLS

To create its estimate of employment the CES uses a survey of establishments selected from the QCEW population of establishments. Using the reported employment from respondents CES estimates a ratio of over-the-month employment change using establishments who respond in both the current and prior months.

$$R_t = \frac{\sum_{i \in MS_{t,t-1}} w_i emp_{i,t}}{\sum_{i \in MS_{t,t-1}} w_i emp_{i,t-1}} \tag{1}$$

Equation 1 describes this calculation, where R_t is the ratio for month t, $MS_{t,t-1}$ is the set of respondents who reported data in months t and t-1, $emp_{i,t}$ is the employment for unit i at time t, and w_i is the sampling weight. MS is also referred to as the matched sample. R_t is applied to the previous months estimate of employment to estimate the current month's employment. Atypical responses are not included in the ratio estimate, but their employment changes are accounted for in a logical manner. These records do not have a strong influence on the CES estimate, so their treatment is not discussed in this paper. For information on CES estimation, refer to the BLS online documentation (BLS, 2009).

To compile its private, total employment the QCEW aggregates the employment for all of the privately owned establishments in its census. The census comes from a collection of Unemployment Insurance filings from each State. The CES provides a current picture of employment situation, while the QCEW is about 6-9 months old.

These two programs have a number of common elements. Since each one measures the same attribute, each program uses the following definition of an employee:

1. Worked at or received pay by the establishment in the specified month.
2. Worked or received pay for the week including the 12th of the month.
3. Include all workers including executives who are not on leave without pay.

The definition is identical on each program's form.

3. Seasonal Differences

While the employment given by the CES and QCEW are typically very similar to each other in the long run, each program shows different seasonal behaviors. Chart 1 shows the difference between the QCEW total and CES estimate for each month from 2003

through 2007. Seasonal patterns are clearly visible between September and March. Within that time frame we have the following patterns:

1. Between September and October the QCEW gains fewer employees than the CES.
2. Between October and December the QCEW gains more employment than the CES.
3. Between December and January the QCEW looses more employees than the CES.
4. Between January and March the QCEW gains more employees than the CES.

The December to January difference is the only focus of this study.

4. Response Analysis Survey

A previous RAS conducted in 1994 also investigated reported employment differences between these two BLS programs (Werking et al, 1995). That RAS consisted of a sample of 8,000 establishments from a select group of 10 states. Its survey questions were divided into the following categories: method, timing, content, and complexity.

The reporting ***method*** includes information on the source used for the CES and QCEW reports. These questions inquired about the following topics:

- The Payroll process, who creates the payroll, and if payroll was used in the report.
- The counting of checks in the report (bonus checks could count the same person twice).
- The purging of records during the year. For example, additional employees may remain on the payroll until an accountant clears them off at the end of the year.

The primary ***timing*** issue involves the week for which the respondent provides data. Respondents should provide the pay period that includes the 12th of the month, and they were asked a question regarding their adherence to this definition.

Content issues relate to different types of employees that should be excluded from the report. For example, workers on leave without pay should not be included.

Complexity includes several different issues that may confound the reports. Report complexity includes the following issues:

- Multiple payrolls or pay periods
- The use of an existing report the respondent uses each month, thereby making their response easier to provide.
- Different reporters for the CES and QCEW
- A systems change such as a new payroll provider or payroll software.

The current RAS selected a sample of 3,000 establishments from a select group of CES respondents who satisfied the following conditions:

- Establishments must have reported employment for June, 2006 through March, 2007 for the CES and QCEW.
- Exclude government units, Professional Employee Organizations, Hospitals, and the Educational services industry.

- Establishments must be in the 2005 and 2006 CES samples.
- CES respondents must be unique. Specifically, each respondent must provide data for only one establishment in the CES sample.

The RAS sample was designed to study specific reporting differences—including those described in section 3. This survey resulted in 1,840 respondents who filled out the questionnaire. To make certain that the RAS respondents maintained the same seasonal patterns observed between the CES and QCEW, ratios were calculated using equation 1 in which the matched sample could only be a subset of the RAS respondents. Chart 2 shows the different ratios calculated with the RAS respondents. The red bars use CES data, and the blue bars use QCEW data. Chart 3 shows the same information using the full set of CES respondents. Except for a few months, the RAS respondents exhibit the same seasonal differences, though those differences are more dramatic in the RAS respondents.

The recent RAS asked questions that were similar to those asked in the 1994 RAS. The current RAS contacted the CES respondent and asked two sets of questions. The first section asked questions related to the CES report, and the second section asked the same questions about the QCEW report. Only 947 of the 1,840 RAS respondents answered questions about the QCEW. Many of the QCEW reports are filled out by a different person, and it was seldom feasible to contact the QCEW respondent. This section nonresponse occurred because the QCEW information was supplied by an alternate source such as a corporate headquarters or professional payroll firm. The analysis incorporated this information to help compensate for the lack of response on the QCEW section of the RAS.

5. Classification and Regression Trees

Classification and Regression Trees (CART) are a data mining tool used for modeling and exploring data. They were introduced by Leo Breiman and their properties are more thoroughly developed by Breiman, Freidman, Olshen, and Stone (1984). Tree methods recursively partition a dataset into mutually exclusive groups by using the following three basic steps:

1. Start with full set of data
2. Select a variable and a variable's value to split the data into 2 mutually exclusive groups. Variable and split are decided by minimizing a loss function.
3. Repeat step 2 on each of the resultant groups.

In predictive modeling a fourth step involves fitting a constant model in each of the final groups, but prediction is not the focus of this research. Steps 2 and 3 are continually repeated until some stopping criterion is reached. The stopping criterion is necessary because the procedure will keep partitioning the data and produce a large number of splits that will overfit the data, and the meaning of this complex structure will be difficult to discern and extend beyond the analyzed dataset. The stopping criterion can be an explicit limit on the number of partitions, a minimum number of observations in each group, a limit on the trees complexity, or a combination of these restraints. The tree complexity discussed in Venables and Ripley (2002) is defined by the following equation:

$$Loss_{cp} = Loss + cp * size \qquad (2)$$

where *Loss* is the sum of the loss function over all partitions, *size* is the number of groups and *cp* is the complexity parameter. The loss function for continuous data is typically the sum of squared errors. Further details on this cost-complexity measure are provided in Venables and Ripley. Equation 2 effectively states that a partition is accepted when it decreases the loss function by at least *cp*; otherwise, it accepts no partition. To decide on this *cp* parameter the analyst typically uses a validation set with observations that were not used to grow the tree (either via a holdout set or by use of cross-validation). The loss function on the validation set is measured for various values of the parameter *cp*, and a minimum for the loss function is often available within the set of considered trees (Venaples and Ripley, p. 258). The *cp* closest to that minimum is chosen. This methodology is available in the *rpart* package available in S-plus and R.

An alternative to the CART methodology is Conditional Inference Trees (Hothorn et al, 2006). This methodology follows the same steps outlined for CART, but step 2 is divided into the following two steps:

2.1 Select variable.
2.2 Select a partition within the selected variable

These two steps are performed to provide an unbiased variable selection property. Unbiased variable selection is defined by a tree selecting a particular variable 1/*p* times if there are *p* covariates and no correlation between them and the dependent variable. CART's preferential selection for certain variables is briefly addressed and displayed by Hothorn et al (p, 651-2, 661-2). Conditional inference trees (CIT) use the asymptotic distribution of a permutation test to form a test statistic, and this method is implemented in the R package *party* (Hothorn). This statistic is used for step 2.1, and while any method may be used to decide the optimal partition in step 2.2, the *party* package uses the same test statistic. The association between X_j and Y is measured using a test statistic computed in the following manner:

$$T_j(\kappa_n, w) = \sum_{i=1}^{n} w_i g_j(X_{ji}) h\left(Y_i, (Y_1, Y_2, ..., Y_n)\right)^{\mathrm{T}}$$

$$\mu_j = E\left(T_j(\kappa_n, w) \mid S(\kappa_n, w)\right) = \left(\sum_{i=1}^{n} w_i g_j(X_j)\right) E\left(h \mid S(\kappa_n, w)\right)$$

$$\Sigma_j = V\left(T_j(\kappa_n) \mid S(\kappa_n, w)\right)$$

$$= \frac{w_\cdot}{w_\cdot - 1} V\left(h \mid S(\kappa_n, w)\right) \otimes \left(\sum_i w_i g_j(X_{j,i}) \otimes w_i g_j(X_{j,i})^{\mathrm{T}}\right)$$

$$- \frac{1}{w_\cdot - 1} V\left(h \mid S(\kappa_n, w)\right) \otimes \left(\sum_i w_i g_j(X_{j,i})\right) \otimes \left(\sum_i w_i g_j(X_{j,i})\right)^{\mathrm{T}}$$

$$E\left(h \mid S(\kappa_n, w)\right) = w_\cdot^{-1} \sum_i w_i \left(h\left(Y_i, (Y_1, Y_2, ..., Y_n)\right)\right)$$

$$V\left(h \mid S(\kappa_n, w)\right) = w_\cdot^{-1} \sum_i w_i \left(h\left(Y_i, (Y_1, Y_2, ..., Y_n)\right) - E\left(h \mid S(\kappa_n, w)\right)\right)$$

$$\left(h\left(Y_i, (Y_1, Y_2, ..., Y_n)\right) - E\left(h \mid S(\kappa_n, w)\right)\right)^{\mathrm{T}}$$

where $S(\kappa_n, w)$ is the symmetric group of all permutations of the *n* respondents, $g_j()$ is a non-random transformation of covariate X_j, and $h()$ is a non-random transformation of the response. Hothorn discusses several different types of transformations for $g()$ and $h()$ (p.

657-8). All of the covariates considered in this analysis were categorical with K_j levels for covariate j, so $g_{ji}(k) = e_{K_j}(k)$, the unit vector of length K with k^{th} element equal to one. The dependent variable in this study is continuous, so the transformation h is the identity or rank transformations.

For all permutation symmetric groups Strasser and Weber (p. 5) state that for class k

$$\left(t_{j,k} - \mu_{j,k}\right) \sim N\left(0, \Sigma_{j,kk}\right)$$

For each variable the test statistics are calculated as

$$c_{\max} = \max_{k=1,\ldots,K} \left| \frac{\left(t_k - \mu_k\right)}{\sqrt{\Sigma_{kk}}} \right|$$

or

$$c_{quad} = \left(t - \mu_k\right) \Sigma_{kk}^{+} \left(t - \mu_k\right)^T$$

where Σ_{kk}^{+} is the inverse of the variance matrix. Note that $c_{quad} \sim \chi^2_{K-1}$. Each of these statistics allows the calculations of a p-value. The variable with the lowest p-value is selected, and the partition point is dictated by

$$A^* = \arg\max_{A} c\left(t_{j^*}^{A}, \mu_{j^*}^{A}, \Sigma_{j^*}^{A}\right)$$

Recall that step 2.1 of the CIT algorithm is variable selection. Using c_{quad} and a rank transformation this procedure performs a Kruskal-Wallis test (Hothorn et al., 2006). The statistic c_{max} uses the standardized test statistics for the K categories of each variable and computes a p-value using the maximum.

The α-level in CITs is a defined confidence level for each test, and after the p-values exceed a certain level of confidence the algorithm stops. Similar to *cp* for CART, α may also be used as a hyperparameter that helps to define the length of the tree. Hothorn mentions several possible ways to use and define this parameter (p. 657). Stopping points used for each tree method in this study are explained in the next section.

Trees have several valuable features used in this analysis. Trees perform variable selection in that they include variables that minimize the loss function (CART) or have a high degree of association with the dependent variable (CIT). Trees can also account for interactions. Interactions exhibit themselves within the tree structure, and an example is provided later.

6. Analylsis

The previous RAS looked at the propensity for an error, with an error defined as a difference between the CES and QCEW reports (Werking, p. 794). To see the effect of each variable the authors reviewed contingency tables to see what variables caused an increased proportion of errors. They were able to identify some factors that attribute to increased errors, but the results did not appear conclusive. For example, while 77% of the respondents with multiple payrolls exhibited an error, 62% of the respondents with a

single payroll also exhibited an error. An analyst would expect multiple payrolls to cause an increase in error, but a significant proportion of respondents with a single payroll still exhibit an error. While that research led to some good insights, it is difficult to make a direct connection to the seasonal errors.

The analysis of the recent RAS looks at how each observation affects the seasonal difference. This influence is gauged by calculating each observation's effect on the estimate of the over-the-month-change for each data source. This effect is measured with each observation's relative residual for each data source. The ratio of the over-the-month change is defined as

$$R_{QCEW} = \frac{\sum_{i=1}^{n} w_{i,comp}\, emp_{i,January,QCEW}}{\sum_{i=1}^{n} w_{i,comp}\, emp_{i,December,QCEW}}$$

where the sum is over the *n* RAS respondents with data in December, 2006 and January, 2007. The relative residual is defined as

$$resid_{i,QCEW} = \frac{w_{i,comp}\left(emp_{i,January,QCEW} - R_{QCEW}\, emp_{i,December,QCEW}\right)}{\sum_{i=1}^{n} w_{i,comp}\, emp_{i,December,QCEW}}$$

where $w_{i,comp} = w_{i,sample} w_{i,RAS}$.

Where the weight $w_{i,comp}$ is the composite weight of the RAS and CES sample weights, and $emp_{i,December,QCEW}$ is the QCEW employment value for December, and the January subscript refers to that specific month. A similar value is computed using the CES data to get $resid_{i,CES}$. A respondent's influence is measured as

$$influence_i = resid_{i,QCEW} - resid_{i,CES}$$

If a respondent's CES data pulls the estimate of the CES ratio higher while its QCEW data pulls the estimate of the QCEW ratio lower then it has more influence on the seasonal difference. Chart 4 gives some examples of observations with different influences. A negative influence implies that an establishment's QCEW data pulls the ratio lower than its CES data.

The denominator in the residual calculation is necessary to adjust for any large discrepancy in the weighted employment for each source. For example, if the weighted QCEW employment is 10% higher than the weighted CES employment for all respondents, then a respondent could show a negative influence though it has the same effect on the QCEW and CES. The negative influence would only occur because of the different employment levels for each data source.

The influence measure is used as the dependent variable in the regression tree methods from section 5. The goal is to find groups of observations that shift or skew negative. A group's influence on the seasonal difference is measured by removing that group from the ratio calculations and measuring how much the seasonal difference changes.

It's possible to explore this data in the same manner as the 1994 RAS—analyzing each variable at the top level and looking for relationships or patterns. Trees effectively perform the same task with the initial partition, but they also look at additional variables after initially selecting the more informative variable, possibly unmasking interactions in the process. The tree effectively strips away observations that are less interesting—symmetric data with influences that balance each other—leaving final groups with a specified set of characteristics, a large seasonal influence, and a smaller number of observations relative to the influence.

Applying the RAS data to CART did not provide any useful results. The data are very noisy and have long tails. The range on the standardized influence is [-10,9]. Leaving the data untransformed produced no splits. Using ranks as the dependent variable provided many splits, but the splits typically partitioned between a small group of observations (less than 40 observations or 2.5% of the respondents) and the rest of the dataset. Only one group provided by CART explained more than 4% of the seasonal difference. That group consisted of 550 respondents (25% of all respondents) and explained 60% of the seasonal difference. The tree did not give a well-defined description of this group's characteristics. At each partition the influential groups typically contained most of the categories in the selected variables, making identification of select attributes difficult.

CITs also gave no splits under an identity transformation, but it was much more informative with a rank transformation using c_{max}. CITs used a nominal confidence level of α=.20. A low level of confidence was used to allow the tree to partition with low significance to find significant interactions with later splits. All of the earlier splits typically occurred with $\alpha < .10$.

Each group but one gives a good set of definable characteristics to which we may ascribe the seasonal difference. The more interpretable groups explained a smaller portion of the seasonal error and will be described later. The indefinable group accounts for the vast majority of the seasonal difference (78%) yet contained only 304 observations (17% of the respondents). Reviewing the characteristics of this group we gain some insight into which possible subgroups make this group so influential.

One outstanding characteristic is that the QCEW and CES are filled out by the same person. This characteristic accounts for 49% of this group, while in the full dataset we see this characteristic in about 33% of the observations. Establishments with this characteristic create a smaller subset of 148 establishments (8% of the full dataset), but they account for 48% of the seasonal difference.

The following list summarizes the main influences in the December to January seasonal error discovered in this analysis, and it includes the previously mentioned group:

- A group in which the CES respondent does not fill out the QCEW report (those completing the QCR would include payroll provider firms, accountants, or corporate headquarters). This group accounts for 18.4% of the seasonal difference but it consists of 2.8% (49) of the respondents.
- Establishments using a Payroll Provider Firm to submit the QCEW report make up a significant portion of the seasonal error. There were two separate groups with this characteristic. The first group accounts for 21% of the seasonal difference but it consists of only 4.4% (77) of all respondents. The other group accounts for 10% of

the seasonal difference but it consists of only 3.5% (66) of the respondents. These two groups are similar to the one mentioned in the previous bullet.

- Establishments using multiple payrolls and/or pay periods, yet they claim that all them are included in the CES report. This group accounts for 16% of the seasonal difference, but it makes up 1.4% (24) of the respondents.
- There's a small subset of respondents in which the same person reports to the CES and the QCEW, and this group accounts for 42% of the seasonal difference yet makes up 8% (148) of the respondents.
- Most employee types do not appear to make a large contribution to the seasonal difference.

These characteristics shed some light on why differences occur. Note that the sum total of the difference explained is greater than 100%. This occurs because there are groups that shift and skew positive, implying that those groups tend to have the opposite influence under investigation.

Aside from the important characteristics mentioned above, there are several other unmentioned factors. Industry played an important role for some of the previously mentioned factors, where the Construction and Professional & Business Service industries tended to have a strong negative influence. Employee types did not play a role in helping to find influential groups. Only those partitions that predominantly did not have particular employee types led to influential groups. This is a potential interaction in the data that would have been difficult to unmask via exploration.

Considering the results and methods of the previous RAS these results may be somewhat unexpected. In the previous RAS employee types such as out of state workers had a significantly higher propensity for error in the CES (Werking et al, pp. 796-8), while the results of this study showed that the treatment of out of State workers did not have a strong influence on the seasonal difference. Perhaps the most counterintuitive result was from respondents in which the same person files both reports. We expect these respondents to provide the same data to both programs, but some of them clearly don't. Perhaps one reason why this group may not have been discovered in previous RAS studies is because the group is so small, while its effect is very large. Considering the number of partitions in the tree, it's unlikely that some of these results would have been detected through exploratory analysis.

7. Summary and Conclusion

The QCEW and CES are two BLS programs that measure of total, private, employment each month. While they measure the same data their seasonal patterns are not the same. In this study I attempted to find some explanation for the December to January difference between the two programs. Previous RAS conducted by BLS explored the data by looking at tables and how some of the characteristics reviewed in this paper increased the propensity for an error between the QCEW and CES. In this paper I reviewed the influences for a particular error, the different changes between December and January. I used tree algorithms to help with variable selection, interaction detection, and identifying groups of observations with a sizeable influence on the seasonal difference. The final results were quite informative and somewhat unexpected, and considering their hidden and unexpected nature it's possible that this information would not have been discovered using the same analyses performed in the previous RAS. When measuring the influence the use of regression trees gives us some additional insight into the data, as it also

identified groups with larger positive influences. Further analysis is necessary to better understand the characteristics of these different groups.

One weakness of tree methods is that variable selection is greedy, as variables are chosen based on minimizing a loss function (sum of squared errors) and then fully included into the model (Hesterberg p. 63-4). Partitions are also dependent on previous partitions, as each one divides the data into mutually exclusive groups. Each partition is optimal, but this does not imply that the entire tree is optimal. While the information gleaned from a tree is beneficial, questions remain regarding whether or not it is the best tree for the data. There are methods that attempt to overcome these shortcomings, such as bootstrap bumping (Tibshirani & Knight, 1997) and Bayesian stochastic search methods (Chipman, 2000), and I hope to explore these methods in future research.

I also hope to extend this analysis to other months to get a better understanding of the seasonal differences that occur throughout the year between these two important BLS programs.

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

Bibliography

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen 1984. Classification and Regression Trees. Boca Raton, FL: CRC Press.

Chipman, Hugh. Edward I. George and Robert E. McCulloch (2000) Bayesian Treed Models. *Machine Learning*. 48, 1-3, 299-320.

Hesterberg, Tim. Nam Hee Choi, Lukas Meier, and Chris Fraley. (2008) Least Angle and ℓ_1 Penalized Regression: A Review. *Statistics Surveys*, 2, 61-93.

Hothorn, Torsten. Kurt Hornik, Ahim Zeileis (2005). Unbiased Recursive Partitioning: A Conditional Framework." *Journal of Computational and Graphical Statistics*, 15, 3, 651-674.

Strasser, Helmut and Christian Weber (1999). On the Asymptotic Theory of Permutation Statistics. *Mathematical Methods of Statistics*, 8, 220-250.

Werking, George S. Richard L. Clayton and Richard J. Rosen. (1995) Studying the Causes of Employment Count Differences Reported in Two BLS Programs. *Proceedings to the Survey Methods Research Section of the Joint Statistical Meetings*. 792-798.

Tibshirani, Robert. Ken Knight (1999). Model Search and Inference by Bootstrap Bumping. *Journal of Computational and Graphical Statistics*, 8, 4, 671-686.

Venables, W.N. and B.D. Ripley (2002). Modern Applied Statistics with S. New York: Springer. 251-266.

Bureau of Labor Statistics. Technical Notes to Establishment Survey Data Published in Employment and Earnings. http://www.bls.gov/web/cestn1.htm

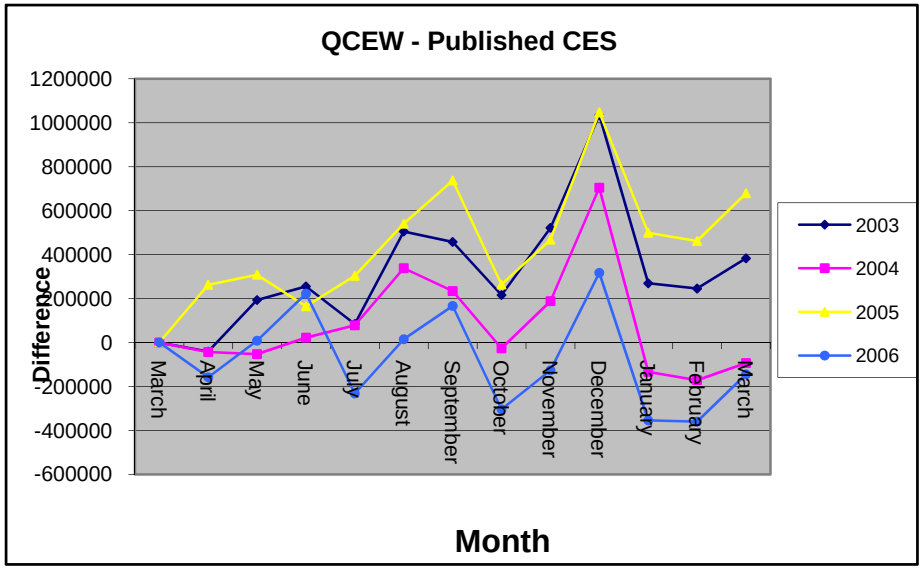

Chart 1: QCEW – Published CES for several years.

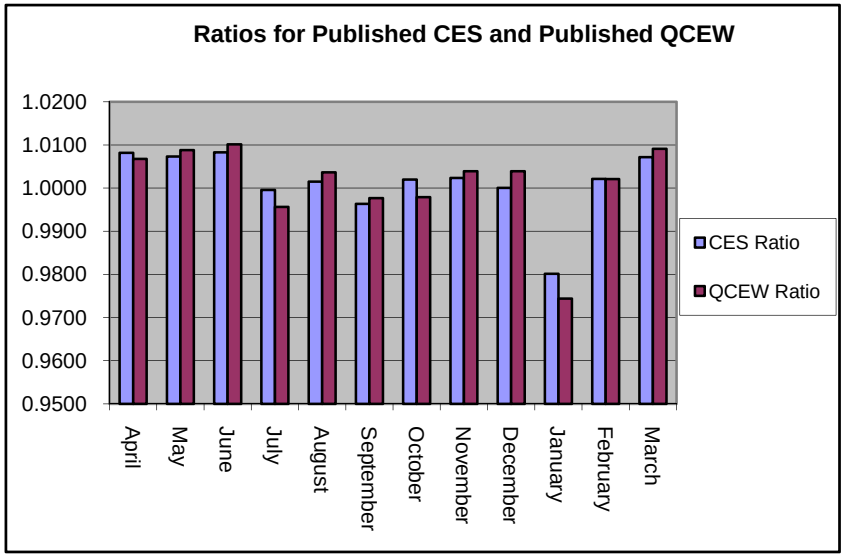

Chart 2: Ratios for Published CES and QCEW

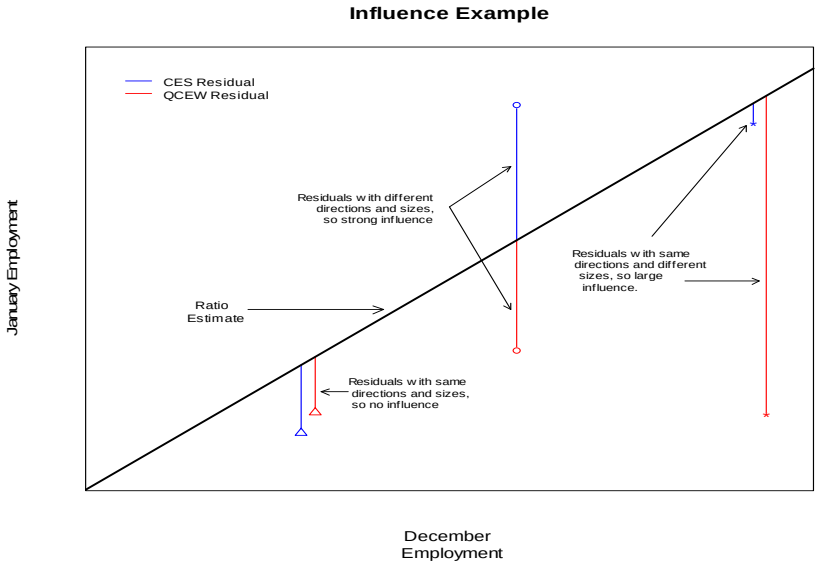

Chart 4: Examples of Influences

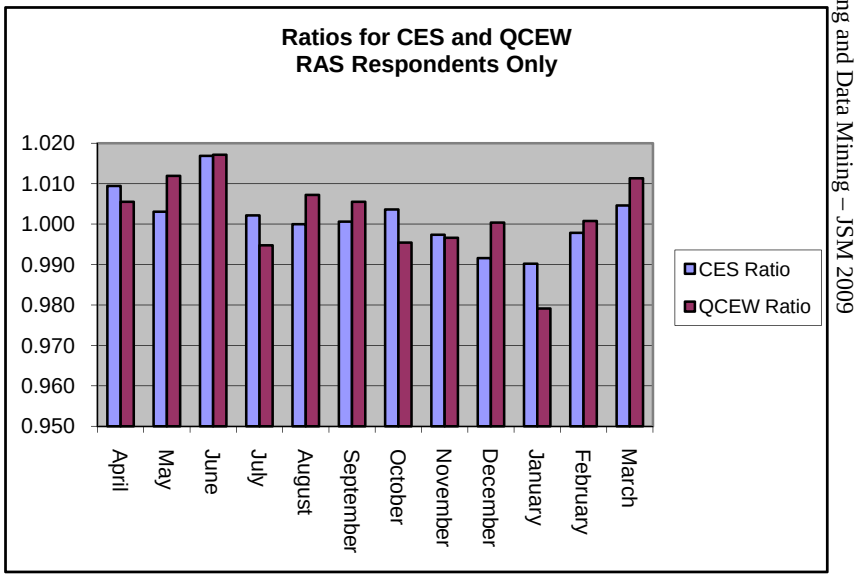

Chart 3: Ratios for RAS respondents using CES and QCEW data