

The Quality vs. Timeliness Tradeoffs in the BLS ES-202 Administrative Statistics

Akbar Sadeghi and Richard Clayton
Bureau of Labor Statistics

Abstract

A primary benefit of using data based on administrative records is comprehensiveness. When timeliness of data is an important factor, however, administrative data are usually less advantageous than sample based data. By the time that administrative records become completely available (usually after meeting a certain deadline), the data are somewhat outdated. Any attempt to expedite the use of such data might result in a loss of accuracy. This paper will investigate the possible trade-off between timeliness and data quality in the case of the Bureau of Labor Statistics' (BLS) Covered Employment and Wages or ES-202. The quarterly ES-202 data are composed of all employers covered by the Unemployment Insurance in the United States and represent a virtual census of employment and wages in the country. The data are compiled by the state employment security agencies and sent to BLS as part of a Federal/State cooperative program. States currently submit their initial micro data files to BLS approximately four months and one week following the close of each calendar quarter. The objective of this research is to determine whether data quality will decrease as a result of receiving these initial files earlier than the current due dates. Six volunteer states submitted their micro data files for each quarter - two weeks prior to the due date, one week prior to the due date, and on the due date. The results for the last three quarters of 2000 showed no significant quality deterioration as a result of moving up the ES-202 due date.

Background

The Bureau of Labor Statistics is the principal statistical agency of the Department of Labor with a mission to acquire and disseminate labor market information (LMI). In execution of its mandate, BLS runs numerous labor force programs each with different LMI objectives. The Covered Employment and Wages Program, commonly known as the ES-202 program, originated with the Social Security Act of 1935 and the initiation of the Federal Unemployment Insurance Act in 1938. The ES-202 program authorized BLS to collect the information needed by the Department of Labor for overseeing the state unemployment insurance programs. Initially, the program covered only private establishments with more than eight workers. The subsequent amendment and other Federal legislation extended the coverage to all firms employing one worker or more and covered public sector workers in the Federal, State and local governments. In 1999, covered employment reported by all sources comprised 97.5 percent of the employees on non-farm payrolls and provided a "virtual census" of employment and wages in the United States.

Under the ES-202 program operating procedures, all employers initially file "Status Determination Reports" with the states unemployment insurance (UI) tax units by which basic business information and classifications are collected. On a quarterly basis, business establishments file "Quarterly Contribution Reports" (QCR) with monthly employment and quarterly wages along with UI tax contributions. In the "Annual Refiling Survey", the status reports are updated for the changes in industry, geographic and ownership codes on a three-year cycle basis with one third of all establishments surveyed each year. The states compile initial status reports, Annual Refiling Survey and Quarterly Contribution Reports, and deliver them to BLS according to a timetable agreed upon in the annual LMI Cooperative Agreement between BLS and the states. Quarterly deliverables from the states are called Enhanced Quarterly Unemployment Insurance (EQUI) files and contain the most comprehensive source of information on employment and wages across industries and geographical locations¹.

ES-202 data have many uses. These data serve as 1) a source of comprehensive macro level employment and wage data by industry and geography, 2) the sampling frame for numerous BLS surveys, 3) a benchmark for critical BLS surveys Current Employment Statistics and Occupational Employment

Statistics and 4) an input to the Bureau of Economic Analysis's National Income and Product Accounts. ES-202 program also annually publishes average employment and wages aggregated by 4-digit industrial classification for the nation, fifty states and the District of Columbia. Data are available to researchers in academia and business community on request from BLS and the states.

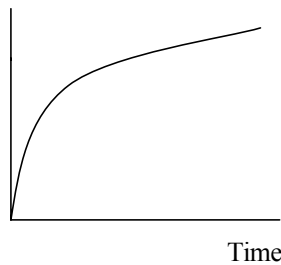
ES-202 represents a prime example of the use of administrative records in data collection. It is comprehensive, cost effective and independent of public cooperation in contrast to the survey-based data gathering. Some of the shortcomings of administrative records, such as the risk of not accurately representing the population of interest, the lack of control on data objects and poor quality of coding, are not present in ES-202 programⁱⁱ. In fact, the ES-202 data has blended well into other survey-based programs at BLS. However, the ES-202 program, like most other data derived from administrative records, is not timely. Employers have a month after the close of the quarter to file UI tax contributions. Late filing and delinquent records are very common. Therefore, administrative data unlike most sample-based data are not readily available. Moreover, the enormity of the number of records makes editing a long and tedious process, which could potentially prolong the availability of completed data. In the case of ES-202, BLS receives the initial microdata files from the states approximately four months and one week after the close of each calendar quarter.

Objective

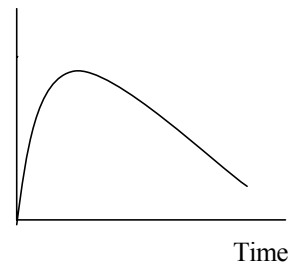
This paper reports the results of an experiment conducted by asking six volunteer states to submit two earlier versions of their EQUI files prior to the due date for the second through the fourth quarter of the year 2000. We asked the volunteer states to submit their single quarter EQUI micro files two weeks and one week prior to the initial due date. It should be noted that there is a final update version that is due a month after the initial version. The due date for update EQUI remained unchanged. We selected one state from each BLS region and attempted to have states that cover a range of sizes and represent both data processing systems. The selected states were Connecticut, West Virginia, Minnesota, Florida, Texas and Alaska. Alaska could not participate in the second quarter 2000 (2000/2) but started with the 2000/3 quarter.

The objective was to determine whether data quality would be lost as a result of moving up the due date by one or two weeks, and to see whether there would be a point when the data quality would deteriorate considerably should the due date continue to be moved up. In pursuing that objective, we implicitly assumed that there is a diminishing marginal return (in terms of data quality) to the amount of time spent on data preparation. Up to a certain point, as more time is spent on editing and data clean-ups total data quality

Total Data Quality



Marginal Data Quality



should increase but marginally at a lower rate. By the same token, if we wish to enhance the timeliness of data by accelerating the due date, less data quality will be lost in the initial unit of time forgone (first two-week in our experiment.)

Methodology

First, we assumed that the counts in the regular due-date version of EQUI are the actual numbers. Therefore, the absolute value of the rate of difference between those numbers and the ones in the early submissions can be used as a measure of data quality. The lower the difference the higher the data quality. A

'no change' is an ideal situation, which means there is no quality loss as a result of moving up the due date. The data quality under this definition is synonymous with accuracy. Second, we selected employment numbers and wages as the main gauge of data quality to represent the entire universe of data in ES-202 files. Finally, we focused on two areas that would be impacted directly and immediately by the early submission of EQUI files: the level of imputations and number of edit failures. Early submittal may cause a higher degree of imputation for delinquent records and a larger number of edit failures due to a shorter time available for editing and data clean-up. As the due date is accelerated, there is a potential risk to data quality since the level of imputed values may rise for less inclusion of reported data. One of our objectives was to measure the change in imputations and their effect on data quality.

Results

EQUI files for the last three quarters of 2000 were delivered as planned in three submittals on due dates for most states. West Virginia transmitted their files several days prior to the due dates in all three quarters, but the interval between submittals was one week like other states. Alaska joined the project in the third quarter for the first time so only two quarters of data were available for this state. Florida and Texas did not send their second submittal in the fourth quarter on time. We could have obtained the files for these two states a few days after the due date, but that would have made the time between the second and third submittal less than one week. To make the numbers more consistent across the states, we decided not to include the second submittal of the fourth quarter for Texas and Florida. For the same reason, the second submittal of the second quarter in Florida was excluded from this analysis. In three instances, all of which occurred in the first submittal, mishaps at the data entry stage produced some unusually large numbers. The errors were not related to the early submission and seemed so large and obvious that could not go undetected under normal delivery. Therefore, we adjusted the numbers for those errors.

In what follows, we will discuss the results in three main areas of concerns: employment and wage data, imputations and edit failures. Employment and wages are the main product of ES-202 program and were used as the main gauge of data quality. Imputations were analyzed to determine if an early submission of EQUI files would impact the level of imputation and if so, what implications that has for the quality of data. Finally, the number of micro edit failures and its impact on data quality are analyzed given the fact that an early submittal leaves less time for editing activities. Table 1 shows the changes in the third month employment and total wages for all volunteer states in all three quarters over different submittals.

- **Total Employment and Wage Data**

In the second quarter 2000, third month employment, a measure of utmost importance for the ES-202 program, showed mixed results with respect to changes from the earliest version to the final one. The difference between the regular due date employment numbers and the first submittal were low in two states (.22 percent in West Virginia and .17 percent in Texas), relatively high in two other states (-2.47 percent in Florida and 1.6 per cent in Connecticut) and acceptable in one state (.82 percent in Minnesota). For a one-week early submittal, the numbers were almost identical in three states and high (-3.59 percent) in Minnesota. All states combined, the total third month employment changed only -.5 percent over the two weekly submittals.

In the third quarter, employment numbers changed slightly over the three submittals. Compared to the numbers in the third submittal (EQUI initial), the change for a two-week early submittal ranged from a low .03 percent in West Virginia to a high .77 percent in Florida. For a one-week early submittal, the change ranged from zero in West Virginia to .61 percent in Texas. All states combined, the total third month employment changed only .47 per cent over two weeks.

In the fourth quarter, the numbers in the first and third submittal were extremely close in all states except Minnesota where the difference was 5.1 percentⁱⁱⁱ. In fact, employment in two states showed no change, and in two other states changed less than one tenth of one percent. For a one-week early submission, all four states for which the data were available, showed close numbers ranging from 0 to .57 per cent. The difference between first and last submittal was .56 percent for all states together.

In the second quarter, except Florida, all states showed a higher third month employment in the final counts compared to the first submittals two weeks earlier, suggesting that an early submittal may underestimate the employment figure. In the third quarter, however, four out of the six states showed lower employment in the final version compared to the first submittal, suggesting the presence of an overestimation as a result of the early submittal. The numbers for the fourth quarter were split with three states showing overestimation and three states showing underestimation. The conflicting results could be interpreted in two ways. It could represent the stochastic and randomness nature of these numbers with no apparent predictable direction. It could also be a result of the seasonal and cyclical factors. The final submittals always include an increase in the number of records with "reported" status that substituted the

Table 1
Changes of Employment and Total Wages over Three Submittals
In percents

	Second Quarter			Third Quarter			Fourth Quarter		
	<i>1st to 2nd submittal</i>	<i>2nd to final submittal</i>	<i>1st to final submittal</i>	<i>1st to 2nd submittal</i>	<i>2nd to final submittal</i>	<i>1st to final submittal</i>	<i>1st to 2nd submittal</i>	<i>2nd to final submittal</i>	<i>1st to final submittal</i>
Employment:									
West Virginia	.23	-.01	.22	.03	.00	.03	.00	.00	.00
Minnesota	4.57	-3.59	.82	.59	.04	.63	4.71	.57	5.28
Texas	.15	.02	.17	-.10	-.61	-.71			.00
Connecticut	1.66	-.06	1.60	-.03	-.02	-.05	-.02	-.03	-.05
Florida			-2.47	-1.18	.42	-.77			-.07
Alaska				.08	-.16	-.08	.00	-.28	-.28
Total Wages:									
West Virginia	.22	.00	.22	.01	.00	.01	.13	.00	.13
Minnesota	.13	.59	.72	.34	-.01	.33	6.30	.43	6.76
Texas	-.02	.08	.05	-.05	.02	-.03			.01
Connecticut	.55	.12	.67	-.36	.04	-.32	-.07	.04	-.02
Florida			.68	.31	.47	.78			.08
Alaska				.16	.0	.16	.0	.0	-.01

previously "imputed" records. Since imputations are mostly based on the past trends and historical values, they may be subject to seasonal and cyclical variations. For instance, in an expansionary period, estimates based on past numbers have a tendency for underestimation, while in a contractionary period the reverse is true. More observations in the future will shed light on this subject.

The pattern of change in employment numbers over three submittals in all three quarters of data does not appear to be linear. The numbers in the early submittals do not increasingly or decreasingly converge to the final due date numbers in a steady and consistent manner. In some instances, total employment increased in the second submittal from the first version but decreased in the final count. The bulk of the change in employment from the earliest to the final version occurred in the second submittal with little change in the last week leading to the regular due date.

The number of records and employment for the "new" and "dropped" establishments changed significantly from one submittal to the next in almost all states and in every quarter. One possible explanation for such fluctuation could be the effect of information about successor and predecessor relations that through time,

becomes available and causes some records to move into or out of new or dropped status. The fluctuating nature of these data can be seen even in the update version of EQUI and will finally settle in the Longitudinal Database where "business closing" and "business opening " are more accurately defined and compiled^{iv}.

The wage numbers followed almost the same pattern as the employment data. However, when compared to the employment figures, the total wages in the early submittals were closer to the regular due date numbers, and there was only one case of an unsatisfactory result. The rate of changes from the early submittals to the final due date version ranged from zero to .78 percent, except in Minnesota where in a two-week early submittal in the fourth quarter, there was a 6.76 per cent difference. The change from the first to the final submittal was mostly distributed between the first and second submittal with less change from the second submittal to the final version. The average weekly wage remained flat with no or minimal change throughout three submittals.

- **Imputation**

In the case of incomplete data or delinquent filings, states impute for the missing data. Later, when the actual data become available, the imputed numbers are replaced by the reported data. To see the impact of imputations on data quality, we needed to know if an early submission of EQUI files would cause a higher level of imputation, and if so how changes in imputation ratios affected data quality. The results showed that the imputation ratios declined across the board for all states over three submittals and in all three quarters. The changes, however, were very small which suggests that in all states, at two weeks prior to the due date, the flow of incoming late unemployment insurance tax filing reached a considerably slow stage.

We ran a statistical regression to see whether variations in data quality can be explained by variations in the imputation ratios. The correlation coefficient turned out very low, showing no significant association between data quality and imputation ratios. The absence of a significant correlation between lower imputations and higher data quality means that the imputation methods must have produced accurate results with estimated numbers compatible with the actual numbers. In other words, a lower imputation ratio simply means that an imputed number has been removed and replaced by the reported data with no impact on total employment and wages.

In order to check the quality of imputation methods, we examined the individual records that were imputed in the first submittal and were converted to the reported status in the subsequent submittals. We did that comparison for the state of Florida and Connecticut in the third quarter and all states in the fourth quarter. With 541 records in the third quarter and 708 records in the fourth quarter, Florida had the largest number of imputed records converted to reported status over two weeks. The total reported employment was very close to the estimated number. However, there were cases where employment was estimated zero in the first submittal but reported with non-zero employment in the final submittal. There were also cases where the opposite happened; a large employment number was imputed in the first submittal but reported zero in the final submittal. Such disparity should be found mostly in the event of either a change in ownership or a change in reporting configuration. A change in reporting configuration happens when a single establishment reports as a multi-establishment employer, or a multi-establishment unit consolidates operation and reports as a single establishment. In such cases, information about successor or predecessor may explain why an establishment with imputed non-zero employment and wages could end up with zero employment and wages. We found the presence of successor information in most zero employment numbers.

The pattern of change in imputed records in Florida was found more or less in other states, especially ones with a larger imputation ratio and a greater number of imputed records converted to reported status. That pattern can be summarized as follows: 1) the total imputed employment and wage numbers were close to the actual reported numbers, 2) individual estimated and actual numbers were close in most observations while overestimated records offset the underestimated records, 3) there were cases of zero employment in both estimated and actual numbers that were converted to non-zero in the subsequent submittal. The new information on ownership change, breakout and consolidation of units that states might have received after

the first submittal could be responsible for the sharp difference between imputed and actual numbers. In General, imputations proved to be neutral in terms of data quality at the macro level, especially at low ratios and small changes.

On the same note, Florida in the second quarter had a 22.4 percent imputation ratio, which is high compared to the national average and Florida's previous quarters' ratios. A high imputation ratio reportedly was caused by a technical problem related to the scanning not delinquency on the part of the employers. That problem worsened in the third quarter and apparently was resolved in the fourth quarter. In the final update version of EQUI for the second quarter, Florida's imputation ratio dropped to 6.8 percent from 22.4 percent. Total employment, meantime, rose to 7,112,271 from 6,953,899. The difference, which is somewhat high, could be a result of the higher reported data that substituted the imputed records, suggesting that a high imputation ratio could have adverse impact on data quality.

- **Edit Failures**

ES-202 data are subject to numerous micro and macro scrutiny initially at the states and finally at BLS. The editing process is predominantly a mechanical one. The computer processes data and assigns edit flags based on some built-in criteria. The state staff review the edit results and correct the erroneous numbers, verify or explain the unusual numbers and check for consistency between Quarterly Contribution Reports and Multiple Worksite Reports. The number of micro edit failures, therefore, shows how many records did not meet pre-set conditions, parameters and tolerances and are grouped in nine levels based on the purpose and severity of the error.

Virtually in all three quarters under study and in all states, the number of micro edit failures steadily declined over the three submittals. The reduction in the total number of edit failures was generally high for two weeks between first and last submittal. However, edit failures related to industrial classification (SIC and NAICS codes) made a large portion of the total edit flags. Adjusting for SIC codes, the remaining number of edit failures showed a decline but at a moderate pace. The number of level five edit messages, which counts the establishments with significant changes in employment and wage over the quarter, changed little or remained unchanged in three submittals. A constant and sometimes sharp decrease in the number of edit failures may indicate that in ES-202 processing flow the last three weeks before the due date are the height of editing activities.

Although the level of edit failures was higher in the early submittals than in the initial due date version in all three quarters under study, they were remarkably in line with previous quarter EQUI initial and the EQUI of the same quarter a year earlier.

To see the impact of changes in the number of edit failures and changes in data quality, we performed a regression analysis. A statistically significant association was not found between the changes in total edit failures (both grand total and non-SIC total) and changes in data quality (measured by the proximity of the earlier submittals to the due date version). However, a linear regression of the changes in level five edit failures on the changes in data quality had a correlation coefficient (R Squared) of .60 and statistically significant coefficients.

Conclusion

We reviewed and analyzed three snapshots of ES-202 data on a one-week-apart basis in three quarters and six states. The results generally showed a lack of significant quality drop as a result of moving up the due dates by one or two weeks. We found a few cases of larger-than-expected deviation from the initial due-date numbers, such as Florida and Connecticut in the first and Minnesota in the fourth quarter. Those cases were not time-related and could not go undetected under any processing system. The absence of a significant quality loss should be attributed to the fact that the experiment with expediting the ES-202 process was done at a late stage in ES-202 timeline, when the marginal improvement on data quality with respect to time was minimal. Generally speaking, a longer time spent on data preparation brings a higher data quality, but marginal return in terms of data quality diminishes as more time are used in the process.

For that reason, speeding up the process by one or two weeks did not cause appreciable quality deterioration. In fact, an early submittal by one week produced numbers generally closer to the actual due date numbers than the same data in a two-week early submission.

Imputation for delinquent records, which showed slight changes over the three submittals, seems to have no apparent implications for data quality, especially at macro level and low ratios. However, new information about the predecessor and successor relation or changes in the status of the establishments as 'continuous', 'new' or 'dropped ' in case of imputed records can create a sharp gap between imputed and actual values. Such information becomes available on a continuous basis over the course of time. The total number of edit failures declined in each submittal but seemed to be neutral in terms of changes to economic data. The level five edit failures, which deals directly with employment and wage numbers, showed rather significant correlation with data quality.

In sum, the observations in this project to date support the idea of moving up the due date by one or two weeks. An extension of this experiment with a larger number of states over fiscal year 2002 will bring more evidence on the plausibility of expediting the processing flow of ES-202 data.

ⁱ "*ES-202 Operating Manual*" U.S. Department of Labor, Bureau of Labor Statistics, July 2000.

ⁱⁱ For more on the uses of administrative records see Ron Provost and Charlene Leggieri "*Expansion of Administrative Records uses at the Census Bureau: A Long-Range Research Plan*" Paper presented at the Meeting of the Federal Committee on Statistical Methodology, Washington D.C. November 1999.

ⁱⁱⁱ Minnesota's numbers in the first submittal of the fourth quarter were underreported due to some records that were erroneously coded as missing. Those records which were subunits of multi establishment employers should have prorated or imputed from the master records. The error was discovered and corrected in the succeeding submittal. Adjusting for this seemingly detectable error, the results for the fourth quarter were very close in all submittals.

^{iv} See Timothy R. Pivetz, Michael A. Searson, James R. Spletzer. "Measuring Job Flows and the Life Cycle of Establishments with BLS Longitudinal Establishments Microdata", *Monthly Labor Review*, Vol. 124, No. 4, April 2001.