Improving Annual Web Collection with Excel Workbooks[1] October 2010

Jennifer M. Harris[1], Michele E. Walker[1], R. Allan Emery[1]
[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC 20212

Abstract
The Current Employment Statistics (CES) Survey, conducted by the U.S. Bureau of Labor Statistics (BLS), annually benchmarks its sample-based employment estimates to independent population controls. The principal source of benchmark data is the Quarterly Census of Employment and Wages, which covers all employees subject to Unemployment Insurance (UI) tax laws. BLS uses several other sources to establish benchmarks for the remaining industries partially covered or exempt from mandatory UI coverage, which accounts for nearly 3 percent of the nonfarm employment total. These data are collected annually via a secure website, which in recent years features an Excel function for uploading and editing data. Collection is still challenging even as the site has evolved to include new features. The impact of the enhancements on data quality, quantity and end user burden will be discussed.

Key Words: Web-based collection, Current Employment Statistics, spreadsheet, benchmark, presumed non-covered employment, data quality

1. Background on the Current Employment Statistics (CES) Program

The Current Employment Statistics (CES)[2] Survey, conducted by the U.S. Bureau of Labor Statistics (BLS) in cooperation with State Employment Security Agencies (SESAs) collects data each month on employment, hours, and earnings from a sample of nonagricultural establishments (including government). The current CES sample includes about 140,000 businesses and government agencies, representing approximately 410,000 individual worksites. From this payroll data, a large number of employment, hours, and earnings series are prepared and published each month with industry and geographic detail.

National CES estimates of employment are one of the first indicators of current economic conditions each month. Preliminary national estimates for a given month are typically published on the first Friday of the following month[3], just three weeks after the reference week (pay period including the 12th of the month). Major data users include the Joint Economic Committee of Congress (JEC), Federal Reserve Board, as well as the financial markets and major media. In addition, CES employment, hours, and earnings data are inputs to other major economic series including Personal Income, Indexes of Industrial Production, and Indexes of Leading and Coincident Economic Indicators.

---

[1] Any opinions expressed in this paper are those of the authors and do not constitute policy of the U.S. Bureau of Labor Statistics.
[2] For more information on the Current Employment Statistics (CES) Program, see http://www.bls.gov/ces/home.htm.
[3] The Employment Situation news release schedule, see http://www.bls.gov/schedule/news_release/empsit.htm.

1.1 Annual Benchmark

Each year, the CES survey realigns its sample-based estimates to incorporate universe counts of employment through a process known as benchmarking. Complete counts of employment, or benchmarks, are derived primarily from the Quarterly Census of Employment and Wages (QCEW)[4]. The QCEW counts are based on Unemployment Insurance (UI) tax reports that nearly all employers are required to file with their State. In this benchmarking process, the difference between the benchmark level and the previously published CES March estimate for each estimating cell is computed. The benchmark determines the final employment levels, while sample movements capture month-to-month trends.

2. Presumed Non-Covered Employment (PNC)

The QCEW accounts for approximately 97% of the CES universe. The remaining 3% consists of employees exempt from State UI tax laws and therefore not counted by the QCEW. Examples include members of religious orders, elected officials and legislators, students working at the college or university they attend and railroad employees. Collectively, BLS refers to this group of employees as presumed non-covered (PNC) employment, and employment counts for this portion of the population must be calculated using alternative sources. With the 2009 benchmark, close to 3.8 million employees were added to the QCEW to complete the CES universe.

There are challenges in determining the size of the population for these groups of employees and in some case defining the group itself. No single source of PNC data exists; therefore BLS uses a number of sources to generate the counts including County Business Patterns and Public Employment data from the U.S. Census Bureau[5], and the Railroad Retirement Board. The source data are published on a lagged basis, so it becomes necessary to extrapolate the "base" number to obtain counts for the current benchmark year. BLS calculates a national PNC employment count by industry for March and then distributes it proportionally to those States where historically PNCs have been reported. States are required to review the March figures and either sign-off on the BLS-calculated values or supply its own updated ones.

3. History of PNC Employment Collection

Traditionally, BLS collected PNC employment data by mail. Each October, BLS would mail a paper form (SO-270) to the SESAs with its estimate of the number of employees by industry that worked during the March reference period[6] but not covered by the Unemployment Insurance tax system. The SO-270 form was divided into three sections based on the ownership code: private sector, state government and local government. The collection form covered 21 employment categories with the private portion containing 13

[4] For more information on the Quarterly Census of Employment and Wages (QCEW) Program, see http://www.bls.gov/cew/home.htm.

[5] For survey definitions and details on the U.S. Census Bureau publications, see http://www.census.gov/econ/cbp/index.html.

[6] Pay period including the 12th of the month. In this case it would be the benchmark month, March.

different industries[7]. The state and local government sections each covered 4 industries (See figure 1). BLS required States to review the BLS-supplied PNC figures and submit a revised count if they had more accurate information about the PNC levels. States were also required to submit the source of their data along with the new figure. After the regional offices reviewed the State submitted data, they would mail the form back to the BLS national office.

Over time, BLS expanded its collection methods to include FAX and email. This did ease respondent burden in some respects but it created another set of processing challenges for BLS. Having multiple transmission methods made it burdensome to track which States had returned the PNC collection form and which method the State had used (paper vs. FAX or email). Secondly, these new collection methods did not eliminate the time devoted to transferring the data to a centralized database. Typographical errors were sometimes an issue with this manual process. In the hopes of streamlining the collection and reducing processing errors, the BLS implemented a web-based form with the 2002 benchmark cycle.

3.1 Early Experiences with Web Collection

The earliest iteration of the PNC collection website, written in htmlSQL and SAS, was very basic in format as well as function (See figure 2). The website displayed an image of the SO-270 collection form and allowed users to input their data online. Web collection did have some benefits, including some improvement to data quality, better organization, timeliness and consistency of the data across States. In addition, it improved the data review process for the BLS regional offices.

3.1.1Data quality checks

There were still several shortcomings with this initial version of the online collection system, including data quality and system performance issues. This early version incorporated some routine edits on the data, confirming that all entries were numeric (or a lettered explanation). It did not however, have any longitudinal edit checks where the system would alert the user of a large change between the current and prior year's submissions. This large over-the-year change could possibly go undetected until the data was received by the BLS national office.

Subsequent versions of the website had the capacity for longitudinal editing and screening. Data are now edited directly on the site before submission, and entries that represent a significant change are flagged for further review[8]. A message detailing which industry, month and value in question will alert States of what they need to address before the submission is considered complete. The State must either enter a corrected value or add a comment explaining the reasoning behind the large change. This alerts the BLS regional office to review this industry before approving the value.

---

[7] With the release of the May 2003 data, the basis for CES industry classification changed from the 1987 Standard Industrial Classification System (SIC) to the 2002 North American Industry Classification System (NAICS). Currently CES uses the 2007 NAICS system.

[8] Flags are based on the percentage change and relative to the historical employment of the industry.

3.1.2 System performance

Unpredictable system performance posed another challenge. The initial version of the system utilized one central SAS dataset as the storage point for all PNC-related data. This was problematic; SAS allows only one user at a time to write to the database. During the collection period all States would be accessing the site attempting to submit data. This caused a significant "bottleneck" which led to extensive processing times and user frustration. This was particularly true during the final week approaching the submission deadline.

To address this "bottleneck" issue, the central SAS dataset was split up into 52[9] individual datasets. Each State was able to access their individual dataset without having to wait in the queue to update the central dataset. Under the new structure, a central database would be updated periodically throughout the day as States completed their submissions. However, this did not completely eliminate performance issues because during updates to the central database, States would be unable to access their individual datasets. Additionally, States publish cross-state Metropolitan Statistical Areas[10] or MSAs comprised of areas that cross state borders. If the PNC datasets were purely „local', a state would not be able to view or use the data from their neighboring states to calculate totals for cross-state MSAs.

These limitations led BLS to change the type of platform the collection website resides on, from a SAS-based to a Sybase-based platform in 2006. Sybase allowed multiple users to update the same data table without interruption to other users. The individual state datasets that were previously used to speed up processing time and minimize wait times were no longer necessary with the new platform. The Sybase database ran on more robust hardware, further enhancing the performance of the site. In 2009, the PNC collection site was migrated to an Oracle platform in order to be in compliance with the new Bureau standard. Under Oracle, BLS has been able to maintain the performance gain realized with the switch to Sybase.

3.2 Expansion of Data Elements

The enhancement of the web-based application presented BLS with the opportunity to expand the detail of the PNC data it collected. The original paper form and primitive web application were limited to statewide level data for March. As the application evolved, States had the ability to submit both monthly data and data at the MSA level (See figure 3). At the national level, this additional data can be used as inputs for the small domain model[11] and for time series reconstruction. The more detailed data is also helpful to the States in their annual benchmark processing. This significantly increased the number of potential observations. For States with a large number of MSAs the possible entries numbered in the thousands. To minimize the extra workload associated with reporting this level of detail, BLS introduced a new feature to prorate the PNC data to the MSA level based on the MSA's percentage of statewide employment.

4. Introduction of Excel Workbooks

[9] This includes the 50 States, the District of Columbia, and Puerto Rico.

[10] For a current list of MSAs see http://www.census.gov/population/www/metroareas/metrodef.html.

[11] The CES Small Domain Model (SDM) is a Weighted Least Squares model used to estimate industries that do not have sufficient sample. For details see http://www.bls.gov/sae/saeguaranteed.htm.

In prior years (2002-05), States had to manually enter their PNC employment on the collection site. In 2006, BLS introduced the Excel workbook option to accommodate the increased number of potential PNC submissions, with the addition of monthly and area level data a few years back. Discussions with State and regional users revealed that many already had PNC data stored in various electronic formats (databases, spreadsheets, and text files etc.). States could take advantage of this pre-existing data and simply cut and paste their data into the pre-formatted spreadsheet and then upload it to the application. This was the most significant enhancement to the collection system to date.

States begin by downloading an Excel spreadsheet that is pre-filled with the State Federal Information Processing Standards (FIPS) codes, MSA codes (state-specific), NAICS industry codes, ownership codes and the BLS-supplied statewide figure for March. This is the same March employment value States would see if they submitted their data manually on the collection site (See figure 4). Updates to the NAICS classification codes or changes to MSA redefinitions are easily incorporated into the spreadsheets before the States begin annual PNC processing. States have the option of either accepting the BLS-supplied PNC values or supplying its own. Users can also enter monthly data, MSA level data and PNC employment for industries not supplied by the BLS. States can upload the data to the central database with the simple click of a button. If changes are necessary, States can download the spreadsheets as often as needed and re-upload with the changes.

4.1 Reducing End User Burden

The primary goal of introducing the Excel upload feature was to make the reporting process easier for State users. Using the number of uploads as a proxy measure of end user burden, we examined how many times users had to upload their data. In some cases, users had to upload their data more than once either due to errors in the data or technical problems with the collection site. Over time, the average number of uploads per state has steadily decreased. In 2006, when the upload was introduced it took states an average of 6 uploads to transmit their data. The following year this dropped to 2 and since then the average number of uploads has hovered around 2. The mode value may be a better indicator of how the upload process has improved or changed over time for most states. From 2007-2009, the mode number of uploads was 1, with 21 states uploading just once in 2009.

The Excel upload feature has not had a significant impact on the timeliness of when states submit their data; however, there is a measurable impact on the states' processing time. We analyzed the system's activity logs to see how long it took states from the time they logged onto the site to the time they marked their status as "complete" and refer to this timeframe as "processing time". On average States who uploaded their data had shorter processing times in comparison to direct web entry States. States using the Excel upload feature took approximately 11 workdays to complete processing, while direct web entry users averaged 16 workdays. Five Excel states were able to complete their processing in just one day, with an additional 5 completing their processing in two days or less. Only 3 direct web entry states were able to complete processing in 2 days or less. While the data can't speak to Excel's impact on actual workload, it does suggest that Excel has streamlined the process and in turn eased the burden on our State users.

4.2 Quantitative Impact of Upload Feature

BLS hoped to see a sustained increase in the number of observations and the level of detail (additional area or monthly data), with the introduction of the Excel upload. There

was a small increase between 2005 and 2006 when the feature was introduced, but the total number of observations peaked in 2006 and has trended downward slightly since then (See table 1). This relatively stable trend can be primarily attributed to the limited scope of the data. States are only required to submit data for the month of March (benchmark month) at the statewide level for a limited number of industries. Therefore, there are a finite number of data elements one state can produce.

4.2.1 Relative maximum

Examining the number of observations States could potentially submit which we refer to as relative maximum, compared to what they choose to report is another way to evaluate the success of the upload feature. For this calculation, BLS assumes complete reporting for all industries where an estimated PNC figure is provided by BLS, multiplied by the number of MSAs, and then by the number of monthly observations. As mentioned earlier, there are instances where the State provides a figure for a known PNC industry while BLS does not have sufficient information to calculate an estimate. In other cases, the State has knowledge of additional industries and provides the industry code and an estimate of PNC employment. This occurs with some frequency, and in some cases will push the percentage of the maximum observations above 100 percent.

For example, if a small state has statewide data and two MSAs, twenty-one industries to report for on a monthly basis, the maximum number of observations would be 756 (See formula below).

$$RM = \sum NI\ (21) \cdot \sum MSAs\ (3) \cdot \sum MO\ (12) = 756$$

Where RM=relative maximum, and
NI=NAICS industry and
MSA=Metropolitan Statistical Area and
MO= number of monthly observations.

We compared each state's potential (relative maximum) with the actual number of data elements the state submitted to create a measure of their reporting completeness. This revealed that Excel upload users tended to submit a greater percentage of their relative maximum. Of the states that submitted more than half of their relative maximum, 88% were Excel upload users. At higher reporting levels this became even more evident. For example, 90% of the states that submitted nearly all their relative maximum- 90% or more- were Excel upload users.

4.2.2 Impact of reporting methods

Further analysis of the composition of reporting methods reveals the impact of the Excel feature. With the initial feature release, Excel uploads accounted for 81% of the total data collected on the site. This reached an all-time high of 97% in 2007. Over the past four years since the feature has been available, an average of 87% of the total PNC data collected has been transmitted through the upload process. The number of States using the upload feature has not increased dramatically but has held constant, with an average of 60% across the four years (2006-2009). Since its introduction more than 82% of States have opted to use the spreadsheet upload at least once. Nearly 40% have used the feature consistently since 2006 and in general these States tend to report more data (relative to their size), compared with States that use direct web entry to submit their data. The remaining 60% of States have opted to use direct web entry most years and utilize the Excel upload sporadically or not at all.

4.3 Qualitative Impact of Upload Feature

The Excel uploads let BLS incorporate instant data validity checks. The collection spreadsheet will not upload with errors. Upon upload, the monthly cells are screened to see if they contain a numeric value or acceptable character value (in cases where PNC data are known but a reliable source is not available). The spreadsheet must contain valid NAICS industry, State FIPS and ownership codes. The upload runs several logic checks including: MSA totals cannot exceed the statewide level, an MSA entry must have a companion statewide figure, and if monthly data is submitted a March figure must be submitted as well (since this is required by BLS). All the failures appear on one screen, specify the type of error and which cells or lines need attention (See figure 5). This significantly speeds up the error review process especially in those cases where the same error appears either across industries or areas. With the manual entry, users must review each error message individually and this can be time-consuming.

5. Final Conclusions

5.1 Advantages of Using Excel

One of the big advantages of using workbooks for uploading data is the user's familiarity with the software. Most if not all State users frequently use Excel in their daily work. Another advantage is the ability of users to easily cut and paste existing data from other sources into the pre-formatted spreadsheet used for the upload. If States collect PNC-related data throughout the year they can simply import the existing data when the PNC processing cycle begins in the early fall. Thirdly, the spreadsheet format is uniform across all states and regions and this makes record keeping and processing more streamlined. Excel spreadsheets can easily be saved for verification or research purposes. States can download these files and easily import them into the state processing system during the CES benchmark.

5.2 Challenges and Changes

Despite the obvious advantages of offering Excel uploads for data transmission, there have been challenges. Some States still prefer to use the direct web entry method since at this time the Excel upload is still optional. Rigid formatting requirements have left some users frustrated with the upload process. Making any changes to the pre-loaded fields in spreadsheet will result in an upload failure. States which have a large number of Metropolitan Statistical Areas may notice significant upload times and in some cases users are unsure if the upload is successful. This leads some users to abandon the upload process and manually enter their data.

BLS has encountered some issues with changes in technology as well. Converting the collection system to an Oracle platform created some unanticipated problems (table sorting and display issues) and required additional testing and programming resources. Recently, the upload process was modified to accommodate the changes to MS 2007 and the. xlsx file format. BLS anticipates making more changes when MS 2010 is introduced.

5.3 Future of PNC Collection

BLS plans to strongly encourage States to use the Excel spreadsheets to transmit their PNC data. Regional Office staff should promote this reporting method and BLS will highlight Excel reporting in the State memorandums. We hope to add additional site logs

with the 2011 collection cycle. Currently, the site tracks page access and when changes occur. BLS hopes to generate an additional log of error messages which users receive when there are either errors in entry (e.g. MSA total is greater than the statewide total) or when a user enters a large over-the-year change. These records would help BLS better understand how the site is used by States, and possibly allow us to create better metrics on data quality.

5.4 Research Summary

Our research leads us to conclude using the Excel uploads has a measurable positive impact on the annual PNC collection process. We've found that states that utilize spreadsheets take less time to process their data compared to direct web entry States, which reduces respondent burden. Excel users tend to submit more of their relative maximum, which translates into more industry and geographic detail. This is a boon for the BLS because it makes our records more complete, providing a better resource for research and future work which may include time series reconstruction and modeling. The greater detail helps the States with their benchmark processing. Overall we feel that the Excel upload has improved PNC web collection, both for our State users and the BLS.

Acknowledgements

We would like to thank Chris Manning and Ken Robertson for their feedback and support on this project.

Figure 1: BLS PNC Collection Form, SO-270 (2001)

Bureau of Labor Statistics
CES Survey State Benchmark Information
U.S. Department of Labor
Presumed Non-covered Employment (PNC)

To: BLS CES Benchmark Section
Through: BLS Regional Commissioner
From: State of North Carolina

Subject: Presumed Non-covered Employment in State in March

Listed below are 21 employment categories presumed non-covered by the ES-202 universe file because these employees are exempt from Unemployment Insurance. Please utilize the BLS developed PNC count from column 3. If a PNC estimate is not appropriate, please provide the alternate PNC source and count, or a lettered explanation (a or b), in columns 4 and 5.

In column 6, please document the reason this action was taken.

Explanations:
a. There is employment in this category in this State, but it is covered under this State's UI laws.
b. There is non-covered employment in this category in this State, but no source of PNC information is available at this time.

2001 PNC FOR **North Carolina**	BLS		STATE		STATE COMMENT/ RECOMMENDATION
	SOURCE	FIGURE	SOURCE	FIGURE	
Total Private 4011, 4013 Other RR	RRB				
4111 Loc/Sub Transit	RRB	0		0	
474 RR Loan Co.	RRB				
63 Insurance	CBP	5121		5121	
6732 Trusts	CBP	0		0	
806 Hospitals	Extrapolation	0		0	
821 Elem. Schools	CBP	0		0	
822 Private College	CBP	14177		14177	
833 Shelter Wk shop	ESA	3011		3011	
835 Child Care	CBP	0			
866 Religious Org.	CBP	78360		78360	

865 and 869 Nonprofit	CBP	221		221	
Other Private	None	0		0	

2001 PNC FOR **North Carolina**	BLS		STATE		STATE COMMENT/ RECOMMENDATION
	SOURCE	FIGURE	SOURCE	FIGURE	
State 822 State College	Public Empl	19157		19157	
806 State Hospital	Trended 1988	0			
91-96 St Government	1987 Census				
State Other	None	0			
Local 822 Local College	Public Empl	0			
806 Local Hospital	Extrapolation	0			
91-96 Local Gov.	SO-270				
Local Other	None	0			

COMMENTS

_____

_____

_____

Regional office should verify the information on this form and return to:

U. S. Department of Labor
Bureau of Labor Statistics
Office of Field Operations
Postal Square Building
Rm 2985
2 Massachusetts Ave NE
Washington, D. C. 20212

_____

SO-270 (revised October 1993)

Figure 2: BLS PNC Web Collection Form, SO-270 (2002)

2002 PNC FOR MINNESOTA	BLS		STATE			STATE COMMENT/ RECOMMENDATION
Private	SOURCE	FIGURE	LAST YEAR	SOURCE	FIGURE	
482112, Short Line RR	RRB	56	128			
485111, Mixed Mode Transit Systems	RRB	0	a			
485113, Bus & Other Motor Vehicle Transit	RRB	0	a			
485999, Other Transit & Ground Passenger Transp.	RRB	0	a			

Figure 3: BLS PNC Web Collection Form, SO-270 (2009)

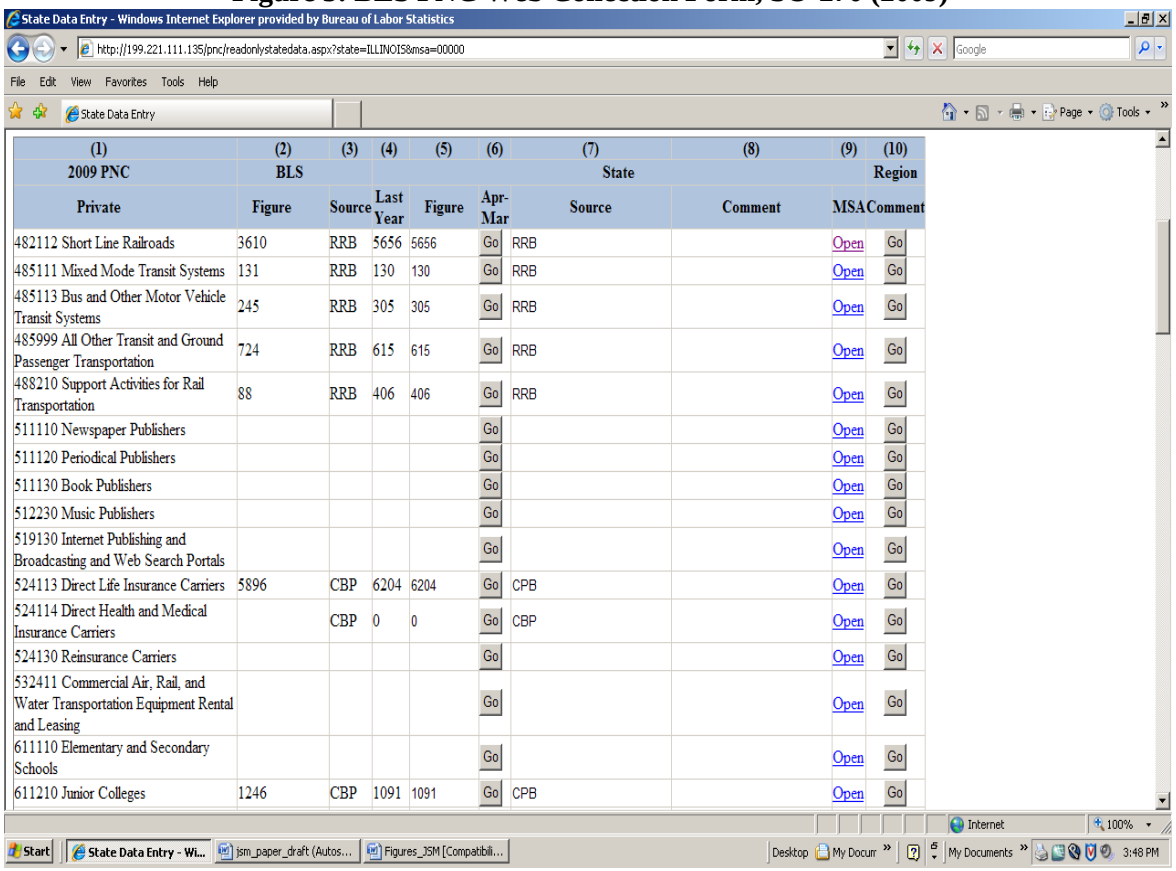
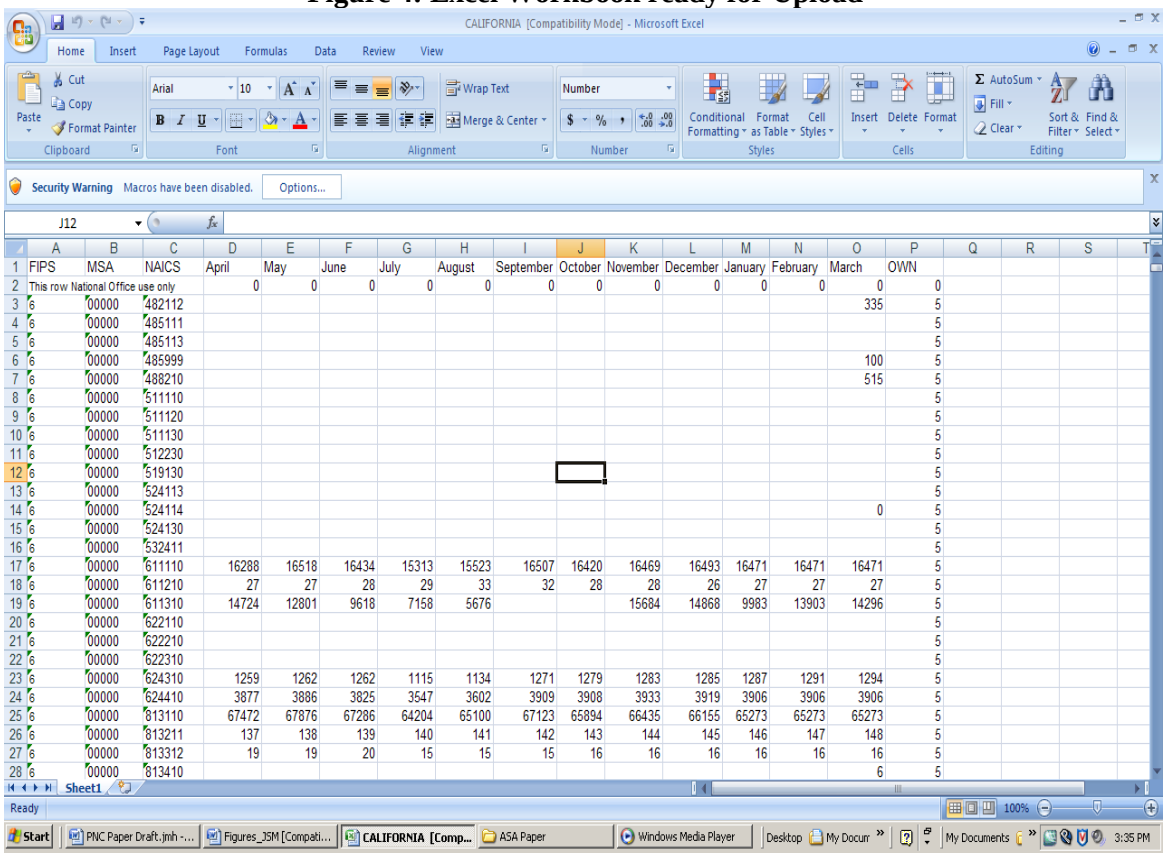


Figure 4: Excel Workbook ready for Upload



CALIFORNIA [Compatibility Mode] - Microsoft Excel

Security Warning Macros have been disabled. Options...

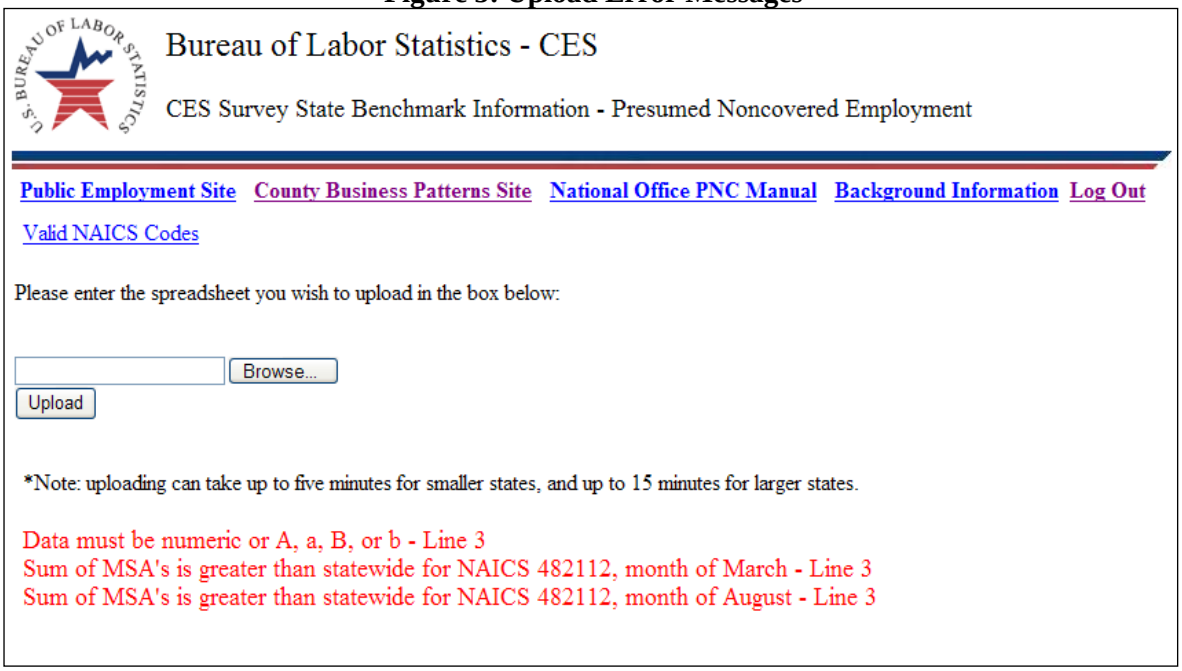
J12

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	FIPS	MSA	NAICS	April	May	June	July	August	September	October	November	December	January	February	March	OWN
2	This row National Office use only			0	0	0	0	0	0	0	0	0	0	0	0	0
3	6	00000	482112												335	5
4	6	00000	485111													5
5	6	00000	485113													5
6	6	00000	485999												100	5
7	6	00000	488210												515	5
8	6	00000	511110													5
9	6	00000	511120													5
10	6	00000	511130													5
11	6	00000	512230													5
12	6	00000	519130													5
13	6	00000	524113													5
14	6	00000	524114												0	5
15	6	00000	524130													5
16	6	00000	532411													5
17	6	00000	611110	16288	16518	16434	15313	15523	16507	16420	16469	16493	16471	16471	16471	5
18	6	00000	611210	27	27	28	29	33	32	28	28	26	27	27	27	5
19	6	00000	611310	14724	12801	9618	7158	5676			15684	14868	9983	13903	14296	5
20	6	00000	622110													5
21	6	00000	622210													5
22	6	00000	622310													5
23	6	00000	624310	1259	1262	1262	1115	1134	1271	1279	1283	1285	1287	1291	1294	5
24	6	00000	624410	3877	3886	3825	3547	3602	3909	3908	3933	3919	3906	3906	3906	5
25	6	00000	813110	67472	67876	67286	64204	65100	67123	65894	66435	66155	65273	65273	65273	5
26	6	00000	813211	137	138	139	140	141	142	143	144	145	146	147	148	5
27	6	00000	813312	19	19	20	15	15	15	16	16	16	16	16	16	5
28	6	00000	813410												6	5

Sheet1

Ready

Figure 5: Upload Error Messages



Bureau of Labor Statistics - CES

CES Survey State Benchmark Information - Presumed Noncovered Employment

Public Employment Site | County Business Patterns Site | National Office PNC Manual | Background Information | Log Out

Valid NAICS Codes

Please enter the spreadsheet you wish to upload in the box below:

Browse...

Upload

*Note: uploading can take up to five minutes for smaller states, and up to 15 minutes for larger states.

Data must be numeric or A, a, B, or b - Line 3
Sum of MSA's is greater than statewide for NAICS 482112, month of March - Line 3
Sum of MSA's is greater than statewide for NAICS 482112, month of August - Line 3

Table 1: PNC Observations by Collection Mode

	Year					
	2004	**2005**	**2006***	**2007**	**2008**	**2009**
Total Observations	35,131	19,345	43,044	29,080	28,564	34,222
Excel States						
Level	N/A	N/A	34,935	27,270	25,189	28,633
Percent	N/A	N/A	81	94	88	84
Direct web-entry States						
Level	35,131	19,345	8,109	1,810	3,375	5,589
Percent	100	100	19	6	12	16

*** Excel upload feature introduced**