Enhancing Web-Based Data Collection using Excel Spreadsheets

Daniel W. Jackson and Michele Eickman

U.S. Bureau of Labor Statistics
2 Massachusetts Avenue, N.E., Room 4860, Washington DC 20212
jackson.dan@bls.gov
eickman.michele@bls.gov

Background on the Current Employment Statistics (CES) Program

The Current Employment Statistics (CES)[1] Survey, conducted by the Bureau of Labor Statistics (BLS), is a monthly panel survey of over 390,000 business establishments. The national CES estimates of employment, hours, and earnings are some of the most timely and sensitive economic indicators published by the federal government and are widely viewed as a key measure of the health of the economy.

Preliminary national estimates for a given reference month are typically published on the first Friday of the following month, after only two and a half weeks of data collection. Because they are one of the earliest indicators of economic conditions each month, CES estimates are used by a wide array of public and private policy makers. Major users include the Federal Reserve Board, Council of Economic Advisors, Joint Economic Committee of Congress, and financial markets in the United States and around the world. CES employment, hours, and earnings data also serve as input to many other economic data series including the National Income and Product Accounts, Indexes of Industrial Production, and Indexes of Leading and Coincident Economic Indicators.

Annual Benchmark

CES estimates undergo annual revisions called “benchmarks”. Each year, the CES sample-based estimates for the previous year are adjusted to universe employment counts derived mostly from the Quarterly Census of Employment and Wages (QCEW)[2]. The QCEW is a nearly complete count of employment for all businesses in the US and is based on administrative records of the Unemployment Insurance tax system that most businesses are required to file with their state. The benchmark adjustment procedure replaces the March sample-based employment estimates with the employment levels from the universe for March of each year. The benchmark therefore determines the final employment levels, while sample movements capture month-to-month trends.

Presumed Non-Covered Employment (PNC)

The QCEW accounts for approximately 97% of the CES universe. The remaining 3% is comprised of workers that are exempt from state UI tax laws and therefore excluded from the QCEW, but are counted by the CES program. Examples include railroad employees, elected officials, students working at the colleges they attend, and members of the clergy. As a group, exempt employees are referred to as presumed non-covered (PNC) employment, and employment counts for them must be derived from alternative sources. In 2006, approximately 3.8 million workers were added to the QCEW figures to complete the CES universe.

The difficulty in making estimates of PNC employment arises from the lack of reliable information on the size of the population for a particular group of non-covered employees, or in some cases, the determination of the group itself. BLS uses a variety of sources to account for the PNC employment, including County Business Patterns and Public Employment Data from the U.S. Census Bureau, information from the Railroad Retirement Board, and in some cases, the States themselves. Since the best source of data often is lagged, it is necessary to extrapolate the "base" number to obtain current estimates. Also, error measures for the derived PNC estimates cannot be computed because information on the population count of the workers is not available.

[1] For more information on the Current Employment Statistics (CES) program, please visit the BLS website at **http://www.bls.gov/ces/home.htm**.

[2] For more information on the Quarterly Census of Employment and Wages (QCEW) program, please visit the BLS website at **http://www.bls.gov/cew/home.htm**.

BLS typically calculates PNC employment on a national level and distributes it proportionally to each State based on the previous year's PNC employment. Each State is responsible for reviewing the BLS-derived figures and either agreeing to them or submitting an alternative value.

Collection of PNC Employment

Traditionally, PNC employment data were collected by mail. Each October, BLS would mail to the State Employment Security Agencies (SESAs) a paper form containing the number of employees by industry who (1) worked the pay period that included March 12 and (2) were not covered by unemployment insurance. The form was divided into three sections based on ownership code: private sector, state government, and local government. The private portion of the collection form covered 13 different North American Industry Classification System (NAICS) industries. The State and local sections each covered 4 different NAICS industries (See figure 1). States were required to review the BLS-provided PNC figures and submit new figures if they had a more accurate count. If they submitted a new figure, they were asked to list the source as well. After their review, the States would mail the form back to the BLS.

Over time, the transmission method expanded to include FAX and e-mail. As a result, it was difficult to track which States had returned the form and in what format the information was submitted (whether paper or electronic). Additionally, hours of staff time was required to transfer the data from the form into a database. Such large amounts of data entry sometimes led to typographical errors.

As a result of these difficulties, BLS developed a web-based form for States to submit their PNC employment. This technology was introduced in late 2002.

Web Collection

The first iteration of the PNC collection website, written in htmSQL and SAS, was basic in its design and had a limited capacity for processing data. The website simply displayed an image of the PNC collection form and permitted users to input their data online (See figure 2). Web collection of PNC employment data had many benefits, including improved data quality and timeliness, better organization and uniformity of the data, and the ability to implement regional oversight. Nevertheless, a few drawbacks were encountered. For example, the initial version of the system incorporated data integrity checks, ensuring that all entries were numeric; however, it lacked longitudinal data edits, so large changes from year-to-year were not immediately evident. Therefore, if California reported twice as much PNC employment in a particular industry compared to the figure reported a year ago, nothing would alert the State of this possible error. This large over-the-year change could possibly go undetected until the data reached the BLS National Office.

Later iterations of the web collection system addressed many of these issues. For example, the next version of the website incorporated longitudinal editing and screening capabilities. Data are now edited directly on the site before submission, and any entries that represent a significant change are flagged for further review.[3] The State must then either enter a corrected value or add a comment to the flagged record indicating the reason behind the large change. This alerts the regional offices and BLS National office of the change.

Performance issues posed another setback. The initial version of the system utilized a central SAS dataset as the storage point for all of the State PNC data; however, one restriction of SAS is that it allows only one user at a time to write to the database. This was problematic because multiple State users would access the website at the same time. This bottleneck occasionally led to extended processing times and user frustration, particularly near the reporting deadline as States rushed to submit their data on time. Also, SAS would sometimes generate HTML too slowly, further impacting performance.

To decrease processing times, the central SAS dataset was split into 52 individual datasets (for the 50 States plus the District of Columbia and Puerto Rico). Each State was able to work on their specific dataset without having to wait in the queue to update the central dataset. The central database was periodically updated throughout the day with the new values submitted by the States. During updates to the central dataset, users were unable to access their individual datasets, because they were used as inputs for updating the central dataset. Despite this processing change, performance issues continued.

These limitations described above led to an eventual change in the platform, from a SAS-based platform to a Sybase platform that can handle a large number of users at once. The advanced locking mechanisms of Sybase allow multiple users to update

[3] A change is considered significant if the increase or decrease is greater than or equal to 50% of the previous level.

the same table with little interference. Multiple datasets, which were once introduced to speed up processing times, were no longer necessary. The database also runs on much more robust hardware, further enhancing the performance of the site.

As the application evolved, so did the ability to collect more detailed PNC data. Originally, the application was limited to collecting data at the statewide level and for March only. In later versions, Metropolitan Statistical Area (MSA)-level data and monthly data were added to the collection instrument (See figure 3). This significantly increased the number of entries States could submit; for example, the number of possible entries for a large state like California went from 54 to more than 21,700. To minimize the workload associated with this increase, a new feature was introduced to the system to prorate PNC data to the MSA-level based on the MSA's percentage of total nonfarm employment at the statewide level. In addition, the enhanced web system produced an extract that was compatible with the State CES processing system. While the additional data was welcomed, the increase put stress on the application and resulted in extended processing times. This led to one of the most significant enhancements to the system: the introduction of Excel spreadsheets as a web-collection method.

Using Excel Spreadsheets in Web Data Collection

When BLS initially began collecting PNC data via the web, users had to enter data manually onto the website. With the increased number of entries that States could submit in later versions of the system, BLS added an Excel spreadsheet feature in 2006 to allow States to upload all of their data at one time in a uniform format. During discussions with states and regional users, BLS learned that many states already had PNC data stored in a variety of different electronic formats (spreadsheets, text files, data bases, etc). The spreadsheet option allowed the states to utilize the data they already had and simply copy and paste from their files to the PNC spreadsheet, further minimizing the amount of manual data entry.

Operationally spreadsheets are easy to use. Respondents download a spreadsheet that is pre-filled with the state Federal Information Processing Standards (FIPS) codes, applicable NAICS codes, and the same employment data as the on-line form (See figure 4). They are able to either accept the BLS-derived PNC figures or update them. They can also enter monthly data if they desire, and add entries for industries where BLS does not provide a statewide figure. Once they have finished, a simple click of a button will upload the spreadsheet to the collection application.

Upon upload, several data quality checks are performed. For example, the system checks for a valid state FIPS code and valid NAICS codes. It also submits the employment data to basic logic and longitudinal edit and screening tests. If no errors are detected, the upload is considered successful and the data are written to the central database. The web form is also updated, and States can then review their data online if they wish. If any problems are detected, the upload is not considered successful. Edit failures are flagged, and the user receives a detailed description of why the upload failed. The user can then return to the spreadsheet, make the necessary changes, and upload the new spreadsheet.

Advantages and Disadvantages of using Excel spreadsheets in Web Collection

The use of spreadsheets was well received by the end users as, they were already familiar with Excel. The spreadsheets decreased respondent burden and cut down on the processing time for most States. Manual data entry was minimized as they were able to utilize pre-existing data to cut and paste into the downloaded spreadsheet. And they were able to save a copy of the final submission for their records, so that they could refer back to their PNC employment figures at any time. The use of spreadsheets also has advantages for BLS. In addition to improved performance of the collection application, the data were received in a uniform format, and the built-in edit and screening checks improved data quality.

There were minor issues with the spreadsheet, particularly with regards to formatting. For example, States would often cut and paste PNC data into the spreadsheet from external sources, and some of these data contained decimals instead of whole numbers. Initially, this caused problems during the upload because the system did not recognize this type of format. A later release of the system, real numbers converted to whole numbers during the upload routine to resolve the issue. Also, while most States reported a decrease in processing time, a few States, particularly those with a large number of MSAs or a large amount of data in their spreadsheet, did report significant upload times of forty-five minutes or more. This was a considerable source of user frustration and caused confusion about whether or not the system was processing the spreadsheet. The problem was related to how the spreadsheet was processed during the upload. Initially, the upload program would check each individual cell in the spreadsheet to see if there had been a change, and after editing the entry, update the database accordingly. In the most recent release of the system, the processing program was adjusted to simply replace the database values with the numbers submitted in the spreadsheet. This cut upload processing times by 75 percent in most cases.

Quantity of PNC Employment Collected

Between 2001 and 2003, BLS collected a total of approximately 600 PNC figures each year from the state users. In 2004, BLS added MSA-level data and monthly data to the online collection form, and the amount of data collected dramatically increased to nearly 36,000 observations. (See figures 5 and 6). In 2005 BLS noted a decrease in the amount of data collected. The application was extremely slow when users were simultaneously submitting PNC data. This led to user frustration and resulted in a smaller number of observations submitted. However, in 2006 the amount of data collected increased again as technology evolved to resolve some of the earlier performance limitations. BLS attributes a large portion of this increase in 2006 to the introduction of the Excel spreadsheet. Positive feedback from users has reinforced this and BLS is testing ways to enhance this collection method.

Conclusion

BLS has had moderate success using Excel spreadsheets to enhance its web-based collection of PNC employment data. In particular, integrating the use of spreadsheets has decreased the user reporting burden while increasing the amount of data collected. Indeed, one of the greatest advantages of offering the Excel spreadsheet functionality is the ability to collect large amounts of data with minimal respondent effort. Compared to manually entering data on the website, uploading and downloading spreadsheets saves significant data entry time.

While not appropriate for all web surveys, the integration of Excel spreadsheets is amenable to surveys that collect simple, alphanumeric data. Excel is a widely used application and therefore familiar to many potential respondents. Once implemented, the spreadsheets will allow for relatively easy expansion of the amount of data collected. BLS will continue to explore ways in which this technology can improve data collection efforts.

Acknowledgements

The authors wish to recognize Allan Emery of BLS for his innovative design and development of the PNC collection application.

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

Bureau of Labor Statistics
CES Survey State Benchmark Information **U.S. Department of Labor**
Presumed Noncovered Employment (PNC)

To: BLS CES Benchmark Section
Through: William D. Pierson, BLS Regional Commissioner
From: Rebecca Rust, State of Florida
Subject: Presumed Noncovered Employment in State in March

Listed below are 21 employment categories presumed noncovered 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 noncovered employment in this category in this State, but no source of PNC information is available at this time.

2001 PNC FOR	BLS		STATE		STATE COMMENT/
FLORIDA	SOURCE	FIGURE	SOURCE	FIGURE	RECOMMENDATION
Total Private 4011, 4013 Other RR	RRB				
4111 Loc/Sub Transit	RRB	0			
474 RR Loan Co.	RRB				
63 Insurance	CBP	7762	CBP	7276	PNC summed from counties
6732 Trusts	CBP	0			
806 Hospitals	Extrapolation	509	CBP	670	PNC summed from counties
821 Elem. Schools	CBP	12915	CBP	11036	PNC summed from counties
822 Private College	CBP	6027	CBP	5409	PNC summed from counties
833 Shelter Wk shop	ESA	4454	CBP	4065	PNC summed from counties
835 Child Care	CBP	767	CBP	4075	PNC summed from counties
866 Religious Org.	CBP	78970	CBP	74367	PNC summed from counties
865 and 869 Nonprofit	CBP	27			PNC summed from counties
Other Private	None	0			
836				226	PNC summed from counties
839				56	Survey conducted in 1989
861				140	PNC summed from counties

2001 PNC FOR	BLS		STATE		STATE COMMENT/
FLORIDA	SOURCE	FIGURE	SOURCE	FIGURE	RECOMMENDATION
State 822 State College	Public Empl	0			
806 State Hospital	Trended 1988	0			
91-96 St Government	1987 Census	0			
State Other	None	0			
Local 822 Local College	Public Empl	14527	CES Data	16164	
806 Local Hospital	Extrapolation	0			
91-96 Local Gov.	SO-270				
Local Other	None	0			

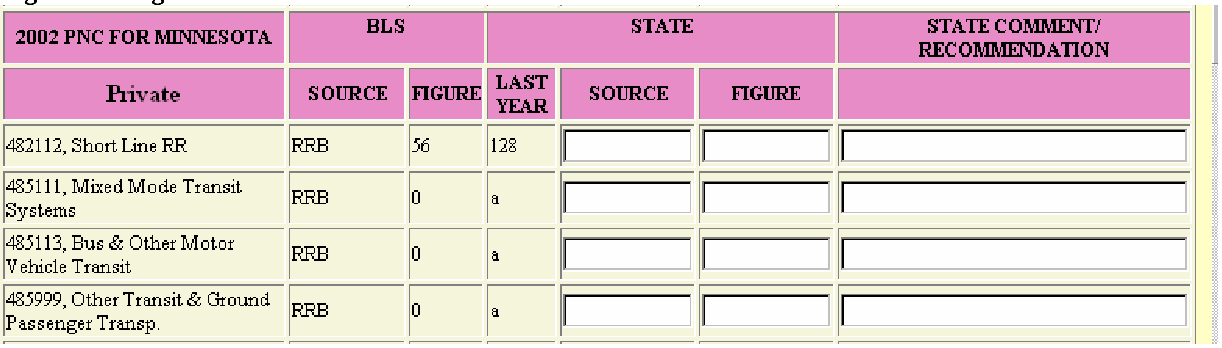
COMMENTS

State Government employment is taken from the Florida State Comptroller's Report and includes noncovered employment.

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

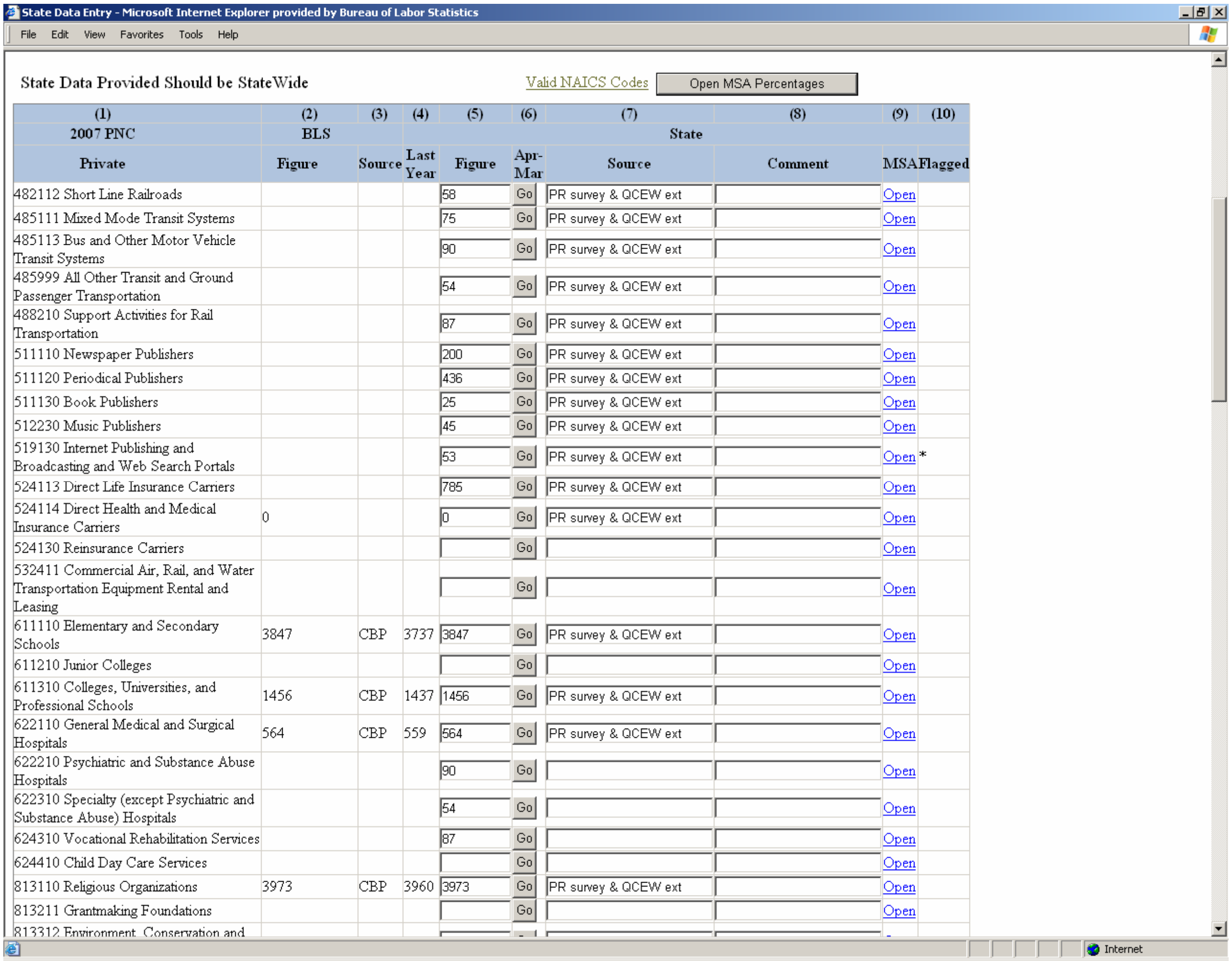
U. S. Department of Labor
Bureau of Labor Statistics
Office of Field Operations
Postal Square Building Rm 2985
2 Massachusetts Avenue N. E.
Washington, D. C. 20212

Figure 2: Original Web-Collection Form



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			

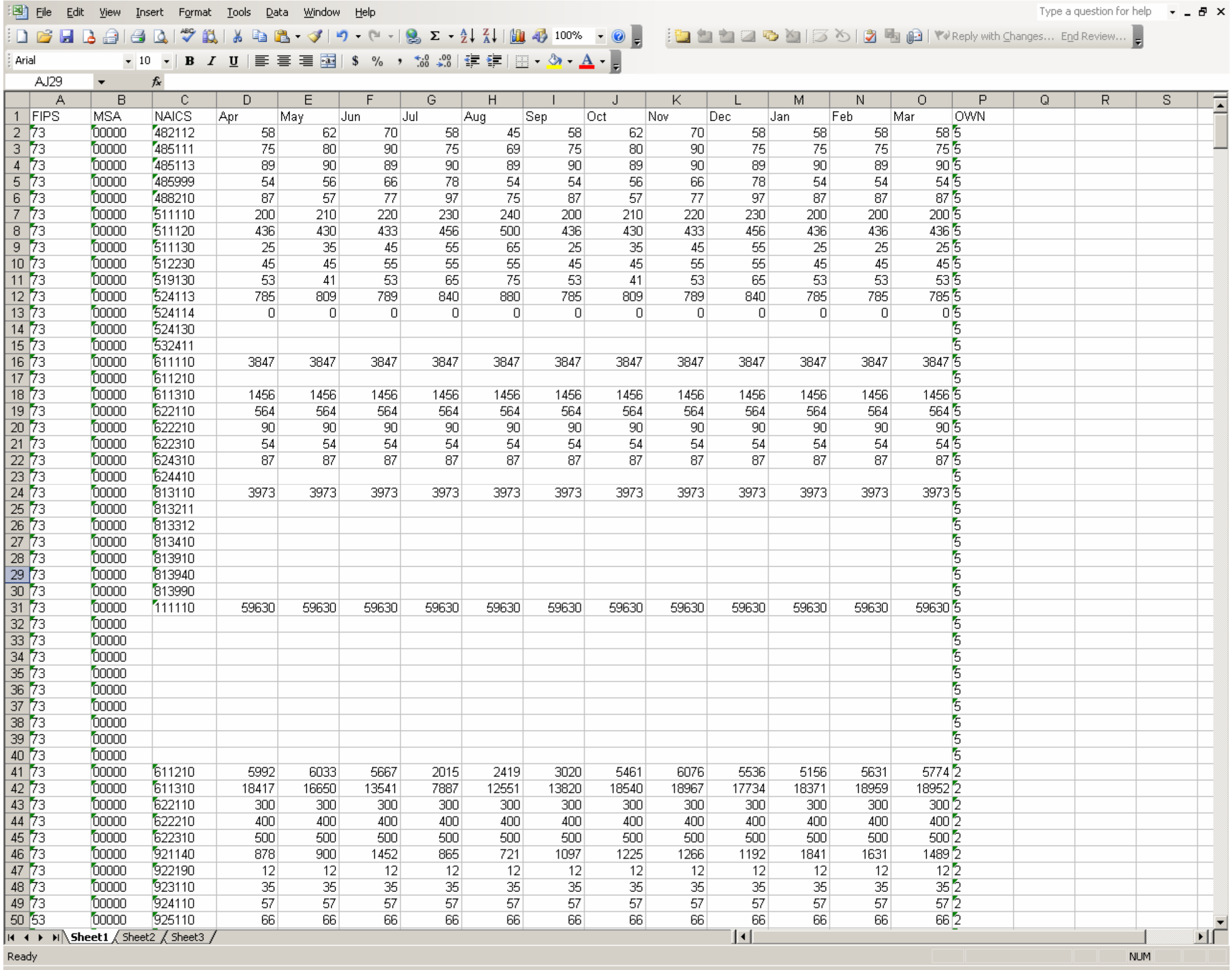
Figure 3: 2007 Web-Collection Form


State Data Entry - Microsoft Internet Explorer provided by Bureau of Labor Statistics
File
Edit
View
Favorites
Tools
Help
Valid NAICS Codes
Open MSA Percentages
Internet

State Data Provided Should be StateWide

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
2007 PNC	BLS		State						
Private	Figure	Source	Last Year	Figure	Apr-Mar	Source	Comment	MSA	Flagged
482112 Short Line Railroads				58	Go	PR survey & QCEW ext		Open	
485111 Mixed Mode Transit Systems				75	Go	PR survey & QCEW ext		Open	
485113 Bus and Other Motor Vehicle Transit Systems				90	Go	PR survey & QCEW ext		Open	
485999 All Other Transit and Ground Passenger Transportation				54	Go	PR survey & QCEW ext		Open	
488210 Support Activities for Rail Transportation				87	Go	PR survey & QCEW ext		Open	
511110 Newspaper Publishers				200	Go	PR survey & QCEW ext		Open	
511120 Periodical Publishers				436	Go	PR survey & QCEW ext		Open	
511130 Book Publishers				25	Go	PR survey & QCEW ext		Open	
512230 Music Publishers				45	Go	PR survey & QCEW ext		Open	
519130 Internet Publishing and Broadcasting and Web Search Portals				53	Go	PR survey & QCEW ext		Open	*
524113 Direct Life Insurance Carriers				785	Go	PR survey & QCEW ext		Open	
524114 Direct Health and Medical Insurance Carriers	0			0	Go	PR survey & QCEW ext		Open	
524130 Reinsurance Carriers					Go			Open	
532411 Commercial Air, Rail, and Water Transportation Equipment Rental and Leasing					Go			Open	
611110 Elementary and Secondary Schools	3847	CBP	3737	3847	Go	PR survey & QCEW ext		Open	
611210 Junior Colleges					Go			Open	
611310 Colleges, Universities, and Professional Schools	1456	CBP	1437	1456	Go	PR survey & QCEW ext		Open	
622110 General Medical and Surgical Hospitals	564	CBP	559	564	Go	PR survey & QCEW ext		Open	
622210 Psychiatric and Substance Abuse Hospitals				90	Go			Open	
622310 Specialty (except Psychiatric and Substance Abuse) Hospitals				54	Go			Open	
624310 Vocational Rehabilitation Services				87	Go			Open	
624410 Child Day Care Services					Go			Open	
813110 Religious Organizations	3973	CBP	3960	3973	Go	PR survey & QCEW ext		Open	
813211 Grantmaking Foundations					Go			Open	
813312 Environment, Conservation and									

Figure 4: Spreadsheet Download



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	FIPS	MSA	NAICS	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	OWN			
2	73	00000	482112	58	62	70	58	45	58	62	70	58	58	58	58	5			
3	73	00000	485111	75	80	90	75	69	75	80	90	75	75	75	75	5			
4	73	00000	485113	89	90	89	90	89	90	89	90	89	90	89	90	5			
5	73	00000	485999	54	56	66	78	54	54	56	66	78	54	54	54	5			
6	73	00000	488210	87	57	77	97	75	87	57	77	97	87	87	87	5			
7	73	00000	511110	200	210	220	230	240	200	210	220	230	200	200	200	5			
8	73	00000	511120	436	430	433	456	500	436	430	433	456	436	436	436	5			
9	73	00000	511130	25	35	45	55	65	25	35	45	55	25	25	25	5			
10	73	00000	512230	45	45	55	55	55	45	45	55	55	45	45	45	5			
11	73	00000	519130	53	41	53	65	75	53	41	53	65	53	53	53	5			
12	73	00000	524113	785	809	789	840	880	785	809	789	840	785	785	785	5			
13	73	00000	524114	0	0	0	0	0	0	0	0	0	0	0	0	5			
14	73	00000	524130													5			
15	73	00000	532411													5			
16	73	00000	611110	3847	3847	3847	3847	3847	3847	3847	3847	3847	3847	3847	3847	5			
17	73	00000	611210													5			
18	73	00000	611310	1456	1456	1456	1456	1456	1456	1456	1456	1456	1456	1456	1456	5			
19	73	00000	622110	564	564	564	564	564	564	564	564	564	564	564	564	5			
20	73	00000	622210	90	90	90	90	90	90	90	90	90	90	90	90	5			
21	73	00000	622310	54	54	54	54	54	54	54	54	54	54	54	54	5			
22	73	00000	624310	87	87	87	87	87	87	87	87	87	87	87	87	5			
23	73	00000	624410													5			
24	73	00000	813110	3973	3973	3973	3973	3973	3973	3973	3973	3973	3973	3973	3973	5			
25	73	00000	813211													5			
26	73	00000	813312													5			
27	73	00000	813410													5			
28	73	00000	813910													5			
29	73	00000	813940													5			
30	73	00000	813990													5			
31	73	00000	111110	59630	59630	59630	59630	59630	59630	59630	59630	59630	59630	59630	59630	5			
32	73	00000														5			
33	73	00000														5			
34	73	00000														5			
35	73	00000														5			
36	73	00000														5			
37	73	00000														5			
38	73	00000														5			
39	73	00000														5			
40	73	00000														5			
41	73	00000	611210	5992	6033	5667	2015	2419	3020	5461	6076	5536	5156	5631	5774	2			
42	73	00000	611310	18417	16650	13541	7887	12551	13820	18540	18967	17734	18371	18959	18952	2			
43	73	00000	622110	300	300	300	300	300	300	300	300	300	300	300	300	2			
44	73	00000	622210	400	400	400	400	400	400	400	400	400	400	400	400	2			
45	73	00000	622310	500	500	500	500	500	500	500	500	500	500	500	500	2			
46	73	00000	921140	878	900	1452	865	721	1097	1225	1266	1192	1841	1631	1489	2			
47	73	00000	922190	12	12	12	12	12	12	12	12	12	12	12	12	2			
48	73	00000	923110	35	35	35	35	35	35	35	35	35	35	35	35	2			
49	73	00000	924110	57	57	57	57	57	57	57	57	57	57	57	57	2			
50	53	00000	925110	66	66	66	66	66	66	66	66	66	66	66	66	2			

Figure 5: PNC Web Collection Statistics

OBSERVATION	2003	2004	2005	2006
Statewide (March)	610	859	794	800
MSA (March)	N/A	5924	2253	4981
Statewide (all months)	610	5215	4437	4926
MSA (all months)	N/A	30458	21937	38893

Figure 6: Monthly Web Collection Statistics

MONTHLY DATA												
2006	Apr.	May	June	July	August	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.
Statewide	374	373	373	377	377	375	375	375	375	376	376	800
MSAs	3072	3064	3056	3090	3090	3088	3088	3088	3088	3094	3094	4981
2005	Apr.	May	June	July	August	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.
Statewide	328	328	328	328	328	329	330	330	330	342	342	794
MSAs	1788	1788	1788	1788	1788	1788	1788	1788	1788	1796	1796	2253
2004	Apr.	May	June	July	August	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.
Statewide	396	396	396	396	396	396	396	396	396	396	396	859
MSAs	2230	2230	2226	2228	2229	2232	2231	2231	2232	2233	2232	5924
2003	Apr.	May	June	July	August	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.
Statewide	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	610
MSAs	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A