The Efficacy of Using Economic Census Data as a Frame Source for PPI *November 2016*

Collin Witt[1]
[1] *U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington DC 20212 U.S.A.*

Abstract
For most industries, the Producer Price Index (PPI) uses for its sampling frame data that originated from the Quarterly Census of Employment and Wages Program of the Bureau of Labor Statistics. Ideally, the proper measure of size for the PPI would be the total revenue of each unit. Because total revenue is not readily available for most of the units, the PPI uses employment size instead. It is believed that employment size and revenue are highly correlated for the manufacturing sector. Considering the fact that the Economic Census (EC) has total shipments and receipts, we investigate the efficacy of using EC data as an alternative frame source for PPI sampling.

Key Words: ***Economic Census, Quarterly Census of Employment and Wages, Producer Price Index, Alternative Frames***

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

1. *Background*

Producer Price Index

The Producer Price Index (PPI) of the Bureau of Labor Statistics (BLS) is a family of indexes that measures the average change over time in the prices received by domestic producers of goods and services. The PPI measures price change from the perspective of the seller. More than 100,000 price quotations per month are organized into three sets of indexes:

i.) FD-ID (Final Demand –Intermediate Demand). The final-demand portion of the FD–ID structure measures price change for commodities sold as personal consumption, as capital investment, to government, or as exports. The intermediate-demand portion of the FD–ID system tracks price change for goods, services, and construction products sold to businesses as inputs to production, excluding capital investment.

ii.) Commodity Indexes. The commodity structure organizes products by similarity of end use or material composition.

iii.) Industry Indexes. The entire output of various industries is sampled to derive price indexes for the net output of industries and their products. PPIs for the net output of industries and their products are grouped according to the North American Industry Classification System (NAICS).

The Longitudinal Database

The PPI uses the BLS Longitudinal Database (LDB), which comes from the QCEW (Quarterly Census of Employment and Wages), as the source of frame information for most of the industries sampled. The LDB contains U.S. business records representing all U.S. non-farm industries, with the exception of some sole proprietors. The LDB consists of all covered employers under the Unemployment Insurance (UI) Tax System. The Employer Identification Number (EIN) is used to cluster establishments, resulting in a single point of contact.

The six-digit NAICS industries are sampled using a two-stage design. First-stage sample units are selected in the Washington office from a list of establishments and clusters of establishments whose primary production is thought to be in a given six-digit NAICS industry. The final or second-stage sample units are then selected during data collection at the location of the sampled establishment. The second-stage units are unique items, products, or services, for which the respondent is to report prices on those selected monthly for 5-7 years.

The first-stage sample units are selected systematically with probability proportional to a measure of size. The measure of size is usually employment when the LDB is used as the frame source for sampling. Employment, which is collected directly from a sampled unit and is used in the weight of items in index calculation, is thought to be highly correlated with revenue. The second-stage sample units are selected by a disaggregation method used in the field at the location of the establishment selected in the first stage.

The PPI already uses aggregated Economic Census data to help determine publication goals, cells for index construction, and index weights. Therefore, utilizing detailed establishment-level data, which includes product-level detail and appropriate revenue information such as the value of product shipments, would seem to be a natural extension for the use of Census data in constructing the PPI frame.

The Economic Census

The Economic Census (EC), is the U.S. federal government's official five-year measure of American business and the economy. It is conducted by the U.S. Census Bureau, and a response is required by law. Forms go out to nearly four million businesses, including large, medium and small companies representing all U.S. locations and industries. Respondents are asked to provide a range of operational and performance data for their companies. Trade associations, chambers of commerce, and businesses use information from the Economic Census for economic development, business decisions, and strategic planning purposes.

Census Database

A single-unit enterprise's primary identifier is its Employer Identification Number (EIN). The Internal Revenue Service (IRS) issues the EIN, and the enterprise uses it as an identifier to report its payroll taxes. All employer enterprises are required to have at least one EIN, and only one enterprise can use a given EIN. Because a single-unit enterprise has only one establishment, there is a one-to-one relationship between the enterprise and the EIN. Thus, the enterprise, the EIN, and the establishment all reference the same physical location, and all three terms can be used interchangeably and unambiguously.

For multi-unit enterprises, however, a different structure connects the enterprise with its establishments via the EIN. A multiunit enterprise consists of at least two establishments. Each enterprise is associated with at least one EIN, and only one enterprise can use a given EIN. However, a multiunit enterprise may have several EINs. Similarly, there is a one-to-many relationship between EINs and establishments. Each EIN can be associated with many establishments, but each establishment is associated with only one EIN. Because of the possibility of one-to-many relationships, we must distinguish between the enterprise, its EINs, and its establishments.

The Census of Manufacturers

The Census of Manufacturers (CMF) is a primary subset of the Economic Census that is composed of NAICS two-digit Sectors 31-33. The CMF includes all employer manufacturing establishments in the U.S. The purposes of the CMF is to provide periodic and comprehensive statistics about manufacturing establishments' activities and production. Title 13 of the United States Code establishes the Economic Census and provides for mandatory responses.

The CMF collects information from single-unit establishment firms and multi-unit establishment firms by means of either a short-form, a long-form, or Federal income tax records. The short-form report collects basic data from establishments on their: kind of business, physical location, type of ownership, operational status, total revenue, employment, and payroll. The long-form report collects additional information from establishments on their: inventories, assets, capital expenditures, identification and cost of materials consumed, and the quantity and value of shipments. The long-form reports are used for about 11,000 products (based on the 2007 Economic Census).

2.) The Efficacy of Census Data in the PPI

Correlations

The PPI has long believed that employment size and revenue are highly correlated for the manufacturing sector. In this study we first wish to establish that there is a strong positive linear correlation between EC Employment and Shipments/Receipts. To this end, we first investigate NAICS Sectors 31-33 (CMF) correlations at the more detailed three-digit level. Three-digit level correlations were chosen because two-digit level data were too aggregated and for data with more than three digits there may be a shortage of observations to establish the linear correlation between the two variables.

i.) Methodology

EC 2007 CMF data were matched/linked between tables by EMPUNIT_ID. An EMPUNIT_ID is an economic unit (i.e., establishment), usually at a single physical location, where business is conducted or where services or industrial operations are performed. It is important to note that these data are contained only in multi-unit enterprises. Pearson correlation coefficients were then used to establish the linear correlation strength between employment and revenue.

ii.) Summary

All the three-digit NAICS Sectors 31-33 (CMF) had very strong positive linear correlations and significant p-values. Therefore, we can reasonably conclude that there is a strong positive linear correlation between EC Sales/Receipts and Employment for the year 2007, at least at the CMF three-digit level (*See Appendix A*).

Product Code Stability

One use of the Economic Census would be in preselecting the number of item quotes of a certain product ahead of time. If we could show some stability of products over time, then it might be worth our time to preselect items before disaggregation. A possible argument against using Census product code data in the PPI sampling process is the length of time between when the detailed EC data become available and when a particular industry is sampled in the PPI. However, if we can establish that the product codes are relatively stable for an industry over time regardless of the source, i.e., either the Economic Census or the annual supplement data such as the ASM (Annual Survey of Manufactures), then we may be able to alleviate some of that concern. The ASM provides sample estimates of statistics for all manufacturing establishments with one or more paid employee. It is conducted annually except for the years when the EC is conducted, and it has approximately 50,000 establishments. The CMF is the universe from which the ASM sample is selected. A new sample is selected every five years, following the Economic Census, and is supplemented annually with company births.

i.) Methodology

Product codes from Census data for the years 2007, 2009, 2010, 2011, and 2012 were analyzed. The 2008 ASM would be drawn from the 2002 CMF. The NAICS sectors were limited to the CMF (Sectors 31-33). For the years 2007 and 2012, product codes were aggregated from 10-digits to 7-digits, and the product values were then summed for each 6-digit NAICS. This was done to compare the years 2009, 2010, and 2011, because the ASM data are collected at the product class-level (7-digits). The product values for the years 2009, 2010, and 2011 were summed by NAICS. To get a final table, the data were then matched by NAICS and Product Class.

ii.) Summary

Overall product class stability in terms of revenue is dependent on industry. The Mean Average Deviation (MAD) was used to classify the stability of products within an industry. For this study, we considered a MAD of greater than .04 for the top two or more products by rank within an industry to not be stable.

For example, manufacturing industries such as Petroleum Refineries, Soft Drink Manufacturing, Automobile Manufacturing, and Petrochemical Manufacturing appear to have very stable product class stability over time.

On the other hand, there are industries whose product class stability seems to vary somewhat over time. Examples of these are Pharmaceutical Preparation Manufacturing, All Other Basic Organic Chemical Manufacturing, and Electronic Computer Manufacturing.

There was some concern that using data from the ASM, which is from a sample as opposed to a census, would cause the product classes to show more volatility. However, this did not

seem to be the case. The majority of the industries studied appear to have relatively stable product classes over time, regardless of the source. Given that product class definitions at the seven-digit level are fairly broad, it makes sense that product class stability would emerge for the majority of industries.

We also looked at PPI weights for product cells collected from the Industry Specific Disaggregation Worksheet (ISDWS) that had been compared to product revenues from the EC data. The purpose of the ISDWS is to allow the Field Economists to select product categories for the first round of Disaggregation and to collect weighting information for those categories. Whereas PPI creates its own product cells, most of those cells were created based on Economic Census data. In matching the product classes from Census data to product cells from the PPI, we attempted to line up the reference period of the sample with the appropriate Census source. We found that in many cases the results varied greatly by industry. The findings also indicated that the PPI could possibly obtain a more representative sample by pre-specifying the number of items for each cell on the ISDWS for this industry.

Efficacy Issues

Ideally, the proper measure of size for the PPI to use would be the total revenue of each unit. Because total revenue is not readily available for most of the units, the PPI uses employment size instead. In the previous section we showed that employment size and revenue are highly correlated for the manufacturing sector, at least at the three-digit level. Considering the fact that the EC has total shipments and receipts, we investigated the efficacy issues of matching the EC to the LDB.

We matched/linked 2007 EC data to 2007 LDB data by EIN and ZIP Code, and we also tried EIN, NAICS, and ZIP Code. However, we found that we could not use EC data as an alternative sampling frame source for the following reasons,

1.) Time Lag - The time between the end of the reference period of the EC and its availability to BLS is approximately 23 months for short-form data and 27 months for long-form data. For example, the 2012 Economic Census data were collected in 2013 and edited in 2014, followed by the short-form microdata being released in November 2014. The detailed long-form microdata were not released by Census until March 2015. Therefore, Shipments & Receipts from the EC take about three years to obtain, whereas QCEW data has about a nine-month lag.

2.) Access Restrictions - IRS data were collected under Title 26 of the US Code. The single-unit establishment data are commingled with Title 26 data; therefore, BLS was not given access to any of the single-unit data. Under the current MOU (Memorandum of Understanding) multi-unit data cover only about 28% of the establishments and approximately 62% of the total employment for the EC.

3.) Classification Discrepancies - There is no central government agency that determines NAICS classification. Therefore, BLS and Census can classify an establishment differently.

4.) Matching/Linking data - There is no unique identifier to match or link data between the LDB and the EC.

Other Comparison Studies

There has been a long-running effort to understand the differences between the two business establishment lists at BLS and the Census Bureau. Both agencies publish industry statistics on the number of establishments, employment, and payroll. When there are significant differences across published estimates for similar economic concepts, many questions are raised by users of the data. Becker *et al* (2005) reported on the discrepancies that resulted from comparing published aggregated data from the two sources. Elvery *et al* (2006) was one of the first published studies that delved into the microdata of the two sources. The authors matched the microdata from the two agencies using the EIN. The EIN is the only ID available on both sources. An EIN-level observation can represent one or more establishments. This fact and the fact that there are many different uses of EINs causes difficulty in matching. Elvery *et al* (2006) additionally found that EINs that represent only a single establishment had higher match rates than EINs that represent multiple establishments. This obviously makes sense, but it also highlights that one of the problems of this study is that we only have access to the multi-unit establishments.

Fixler and Landefeld (2006) summarized the importance of having consistent data concepts for Bureau of Economic Analysis (BEA) data products that rely on both BLS and Census data. They strongly recommended increased data sharing between BLS and Census as a way of rectifying some of these differences. This external as well as internal pressure to better explain discrepancies between the two sources eventually led to the current MOU between BLS and Census.

Prior to the signing of the MOU between BLS and Census, FitzGerald *et al* (2011) examined the microdata between the LDB and the 2007 Census of Manufacturers (CMF) and Census of Wholesale Traders (CWH) as part of an IPP-PPI Frame Comparison Study. This was the first study done within OPLC that was examined microdata between the two sources. BLS staff went to the Census Data Research Center in Suitland to conduct this analysis. This team had access to both single and multiple-unit establishments that are on the product code tables. However, the team did not have access to NAICS codes, because they were on a data set that was a mix of collected and administrative IRS data. Therefore, the team had to derive the NAICS code from the product codes. This could partially explain the difficulty this team had with matching EINs by industry. Ultimately, they concluded that PPI could not use the Economic Census data as a frame source, because it is available only once every five years.

3.) Conclusion

Based on our current study and previous studies about the use of the Economic Census data, there are several obstacles that prevent the PPI from using these data as a primary sampling frame source. Some of these are a time lag that introduces accuracy issues, the NAICS classification discrepancies, and the matching and linking of data between frame sources. So while the efficacy of using the Economic Census as a primary sampling frame source is not supported, it is possible that the EC could be used for frame refinement and in the selection of second-stage sampling units. Furthermore, Economic Census data are already being used to help determine PPI publication goals and cells for index construction,

so it may also be useful in helping to determine the NAICS code associated with some multi-unit establishments on both the QCEW and the Economic Census. We also see some hope of using the EC data for the preselection of products. Consideration is currently being given to having a team investigate the possibility of using the EC data as a frame refinement source and to determine if we would gain anything from preselecting items.

References

Becker, Randy, Joel Elvery, Lucia Foster, C.J. Krizan, Sang Nguyen, and David Talan (2005). "A Comparison of the Business Registers Used By the Bureau of Labor Statistics and the Bureau of the Census." Paper presented at 2005 Joint Statistical Meetings.

Elvery, Joel, Lucia Foster, C.J. Krizan, and David Talan (2006). "Preliminary Micro Data Results from the Business List Comparison Project." Paper presented at 2006 Joint Statistical Meetings.

Fairman, Kristin, Lucia Foster, C.J. Krizan, and Ian Rucker (2008). "An Analysis of Key Differences in Micro Data: Results from the Business List Comparison Project." Paper presented at 2008 Joint Statistical Meetings.

FitzGerald, Jenny, James Himelein, Rod Meaney, Mike Sibel, and Dave Slack (2011). "Phase III of the IPP-PPI Frame Comparison Study: Overlap between PPI's Longitudinal Database (LDB) and the US Census Bureau's 2007 Census of Manufacturers and Census of Wholesale Traders." BLS internal memorandum dated December 16, 2011.

Fixler, Dennis and J. Steven Landefeld (2006). "The Importance of Data Sharing to Consistent Macroeconomic Statistics," Kuebler, Caryn and Christopher Mackie, eds. in *Improving Business Statistics Through Interagency Data Sharing: Summary of a Workshop*. Washington DC: National Academy of Sciences.

Teresa Hesley, Steven Paben, Andy Sadler, David Slack, and Collin Witt (2015). "Analysis of Census Microdata for Potential Uses in the PPI Sampling Process". Internal BLS Paper.

Appendix A.

3-digit NAICS	NAICS Title	R^2
311	Food Manu.	63%
312	Beverage & Tobacco Product Manu.	78%
313	Textile Mills	68%
314	Textile Product Mills	79%
315	Apparel Manu.	53%
316	Leather & Allied Product Manu.	59%
321	Wood Product Manu.	69%
322	Paper Manu.	77%
323	Printing & Related Support Activities	83%
324	Petroleum & Coal Products Manu.	74%
325	Chemical Manu.	67%
326	Plastics & Rubber Products Manu.	72%
327	Nonmetallic Mineral Product Manu.	62%
331	Primary Metal Manu.	64%
332	Fabricated Metal Product Manu.	73%
333	Machinery Manu.	80%
334	Computer & Electronic Product Manu.	76%
335	Electrical Equipment, Appliance, & Component Manu.	75%
336	Transportation Equipment Manu.	81%
337	Furniture & Related Product Manu.	79%
339	Miscellaneous Manu.	83%