Implementing an Alternative Data Source to Estimate Producer Price Indexes for Selected Financial Services[1]

by

Kathleen Frawley[2], Chief, Finance Information, Real Estate, and Professional Services Team, and Jeffrey Hill, Assistant Commissioner
Producer Price Index, U.S. Bureau of Labor Statistics

Presented at the ASSA Annual Meetings, San Diego, CA, January 5, 2020

Session: Improving Economic Price Statistics through the Use of Alternative Data

Session sponsored by AEA Committee on Economics Statistics

Discussant: Emi Nakamura, University of California, Berkeley

Abstract

The Bureau of Labor Statistics (BLS) aims to measure the average change over time in the selling prices received by domestic producers for their marketed output through its Producer Price Index (PPI). BLS economists have difficulty finding "big" datasets that include the necessary inputs for estimating the change in prices that producers receive for the goods they produce and the services they provide. This paper describes the process BLS used for one area of financial services, where success came after several attempts and still provides opportunities for expansion. BLS is using a large, purchased database for financial services within the U.S. economy. Economists extract thousands of data points per day to compute a weighted average price for use in PPI estimates. BLS is using this source to replace directly-collected data for municipal debt securities dealing, corporate debt securities dealing, and equities securities dealing in the investment banking and securities dealing industry. For corporate bond dealing, BLS increased the number of transactions measured per day by 3,971 percent; for municipal bond dealing, by 6,640 percent. The PPI also blends data from the same database with directly-collected data to escalate merger and acquisition deal values and underwriting transaction values in the investment banking industry. For comparison, we describe BLS' success obtaining large datasets with transaction data for a few respondents in retail trade services. The PPI relies on retail margins to estimate price change for this sector. One company provides a dataset with acquisition and sales prices from which the PPI can compute the margin information for over 30,000 products covering 15 different item groupings. For both experiences, we describe what circumstances led BLS to consider these alternative sources and collection methods for the PPI, the research and decision points, the results, and future plans.

JEL Codes: C43, E31, G12, L84

Keywords: producer price index, price level, equities, fixed income securities, financial services

---

[1]*The authors thank current and former staff at the U.S. Bureau of Labor Statistics for their valuable contributions to the research and subsequent production activities, especially Andrew Baer, Ralph Bradley, Michael Conforti, Bonnie Murphy, Melanie Santiago, Sara Stanley, Bill Thompson, and Jason Wrobel and thank Thesia Garner and David Friedman for discussions, comments, and guidance. This is a working paper at the Bureau of Labor Statistics. The views expressed herein are those of the authors and not necessarily those of the Bureau of Labor Statistics or the U.S. Department of Labor.*

[2] *Corresponding author: Frawley.kathleen@bls.gov, Bureau of Labor Statistics, Producer Price Index Program.*

1. Introduction

This paper presents some research into and actual implementation of two approaches the Bureau of Labor Statistics (BLS) is taking to advance its measures of producer prices by replacing traditional data sources and collection methods with ones that improve representation of the economy. This improved representation comes from the use of "big data." In addition, big data has the potential to improve BLS' management of respondent burden, to achieve cost savings, and to expand detailed coverage of the economy.

BLS measures the average change over time in the prices domestic producers receive for their marketed output through its Producer Price Index (PPI). The PPI captures individual prices from businesses for their goods and services, commonly referred to as "products," on a monthly basis and aggregates them in a variety of ways to provide data users with different perspectives on price change. *Industry* PPIs are organized and published according to the North American Industry Classification System (NAICS). They provide data on how prices for products produced by an industry change and are useful to businesses for comparing their own price change to the nationwide average for their industry. *Commodity* PPIs follow a unique classification scheme. They provide price change data for products regardless of what industry produces them. Businesses often use this data for price adjustment clauses in contracts. The *Final Demand - Intermediate Demand* (FD-ID) series are BLS' "headline" PPI statistics that analyze high-level inflation. Final Demand PPIs measure price change for outputs sold as personal consumption, as capital investment, to government, and as exports. Intermediate Demand PPIs measure price change for outputs sold to businesses as inputs to production, excluding capital investment. Finally, in lieu of conducting a survey of U.S. businesses to estimate change in the costs of their inputs to production, BLS calculates *Net inputs to industry* indexes for selected six-digit NAICS industries, excluding capital investment, labor, and imports based on data it collects for the PPI. This data series helps U.S. businesses compare the change in their input costs to the nationwide average for their industry. (In 2020, BLS will publish an expanded, experimental Net inputs to industry series for most three-digit NAICS industry groups that will include imported inputs by blending data from the BLS U.S. Import Price Index.)

BLS employs a standard methodology for sampling and data collection for the PPI. The methods are organized around the NAICS. For approximately 575 NAICS industries, the PPI constructs a frame that includes all establishments classified into an industry and then selects a subset of sampled establishments that are representative of the industry as a whole. Not all industries are sampled at once; an industry sample remains in the PPI survey for, on average, seven years before it is resampled, although resource constraints have extended the lifespan of industry samples in the past and are likely to extend them further in the future. The lifespan of any particular industry sample is influenced by the nature of the industry combined with funding constraints. For instance, BLS resamples some industries more frequently than others due to the velocity of change in the establishments entering or exiting the industry and/or the velocity of change in an industry's products.

Once BLS selects an industry's sample, because the PPI is a voluntary survey staff contacts each establishment to solicit cooperation and, if cooperative, initiates it into the PPI survey and directly collects pricing data. During this initiation process, a BLS economist selects the transactions for which prices will be tracked from among all of the establishment's revenue-generating activities by following a probability sampling technique called disaggregation. The probability of a transaction's selection is proportionate to its value within the establishment. Disaggregation begins with broad transaction categories and may continue according to additional detail of the transactions until unique transactions are identified. In order to limit the burden on participating establishments, BLS typically requests to track a total of four to eight transactions per month from each sampled establishment. After initiation into the

survey, each month BLS sends respondents at establishments an e-mail asking them to log into a secure system and provide current period prices for each of the selected transactions. These prices are tracked on a monthly basis over several years until the industry's sample is replaced by a new sample or the establishment attrits for other reasons.[3]

BLS price indexes 1) follow the "matched model" where respondents report on exactly the same item (good produced or service provided) over time, and 2) measure constant-quality price change over time. Thus, when a unique item is no longer available, the respondent provides a replacement item, and any quality change between the original and replacement item must be estimated and removed to reflect pure price change.[4,5]

For many service sectors in the U.S. economy, measuring average price change based upon four-to-eight transactions each month by sampled establishments may be less representative of price behavior than desired, so BLS adjusts its methods accordingly and continuously strives for improvement. BLS has new techniques for some of the data it collects for the financial services and retail trade sectors of the economy, but research into these alternative data sources and collection methods does not always yield productive results. This paper describes the process BLS used for one area of financial services, where success came after several attempts and still provides opportunities for expansion. A second example in retail trade describes a different approach, one that is yielding benefits but moving at a slower pace. In addition to legal and budgetary requirements involved with any alternative approach, or use of "big data," it is imperative that PPI receive and be able to process a big dataset in time for the publication of the reference month's indexes, which normally occurs within ten business days after the month ends. We refer to these as operational requirements.

The remainder of the paper is organized as follows: Section 2 provides background on the *Investment banking and securities dealing* industry and BLS' initial methodology for measuring producer price change and the challenges that resulted. Section 3 describes the search for alternative data sources for this industry, and Section 4 describes the improvements made. For comparison, Section 5 describes BLS' success at obtaining large datasets with transaction data for a few PPI respondents in a retail trade industry. Section 6 provides noteworthy information on legal and budgetary hurdles BLS must overcome related to alternative data sources and collection methods. Section 7 concludes with a look at plans for the future.

2. The Investment banking and securities dealing industry

BLS began publication of the PPI for *Investment banking and securities dealing* (the 2012 NAICS labels this as industry 523110) in 2003. Prior to 2003, the industry was combined under a single Standard Industry Classification code with Security brokerage services. The PPI followed standard methodology for sampling and data collection for the first three samples for the industry for most transactions, although as this paper describes, even in the early 2000s it incorporated some alternative data.

2.1 Industry definition

Establishments classified in *Investment banking and securities dealing* are primarily engaged in underwriting, originating, and/or maintaining markets for issues of securities. Investment bankers act as

[3] For more details on sampling and data collection, see "Chapter 14. Producer Prices" in the *BLS Handbook of Methods*," p. 2-3. https://www.bls.gov/opub/hom/pdf/ppi-20111028.pdf

[4] For more details on constant quality, see "Chapter 14. Producer Prices," p. 3-4.

[5] For more details on quality adjustment, see "Quality Adjustment in the Producer Price Index." https://www.bls.gov/ppi/qualityadjustment.pdf

principals (i.e. investors who buy or sell on their own account) in firm-commitment transactions or act as agents in best-effort and standby commitments. This industry also includes establishments acting as principals in buying or selling securities generally on a spread basis, such as securities dealers or stock option dealers.

Firms within this industry derive a large portion of their income from interest, dividends, and capital gains from the securities held in their own accounts. Interest, dividends, and capital gains earned from these investments are not considered output-generating activity and are not in scope for the PPI. These receipts may be referred to as proprietary trading turnover. While some firms may define proprietary trading to include all trading activities, the PPI defines proprietary trading as only trading that is done on behalf of a firm's long-term investment account. If, through trading activity, a firm takes ownership of a security with the intent of reselling it on the behalf of a client, under the PPI definition this activity is not regarded as proprietary trading and is in scope. Specifically excluded from this industry are: 1) establishments primarily engaged in acting as agents (i.e., brokers) in buying or selling securities on a commission or transaction fee basis; they are classified in Industry 523120, *Securities Brokerage*; and 2) investment clubs or individual investors primarily engaged in buying or selling financial contracts (e.g., securities) on their own account; they are classified in Industry 523910, *Miscellaneous Intermediation*.

Securities dealers, which are classified in this industry, may at times be confused with securities brokers, which are classified in NAICS 523120. Brokers facilitate trades between clients and charge commissions. Operating as go-betweens, securities brokers do not take legal ownership of securities and do not assume any trading risk. Conversely, dealers purchase securities for and sell securities from their own inventories, assuming risk in these transactions. Securities dealers earn revenue based on the spread at which they sell and purchase securities. A broker-dealer is allowed to operate in either role, but never as both at the same time.

Table 1 shows the U.S. PPI calculation structure for NAICS 523110, *Investment banking and securities dealing*:

TABLE 1:

Index Code	Index Title
523110	Investment banking and securities dealing
523110P	Primary services
5231102	Dealer transactions
523110201	Dealer transactions, equities
523110202	Dealer transactions, debt securities and all other trading
5231103	Investment banking services
5231104	Other securities dealing services
523110SM	Other receipts

not all indexes are published

2.2 First three NAICS samples

The first sample for the *Investment banking and securities dealing* industry followed the standard PPI methodology for sampling and data collection. For cooperating respondents, BLS economists selected a limited set of transactions by disaggregating among broad service categories that were closely linked to index lines in the table above. Table 2 summarizes these broad categories and maps them to an index line:

TABLE 2:

Broad Service Category	Definition	Examples of Securities Included	Index Line
"Dealer transactions, equity securities"	Includes transactions in which the firm acts as a principal in buying or selling equity securities for the purpose of executing trades.	Stocks and Exchange traded funds	523110201
"Dealer transactions, Treasury securities"	Includes transactions in which the firm acts as a principal in buying or selling US Treasury securities for the purpose of executing trades.	Treasury bills, notes, and bonds; any other security issued by the US Treasury	523110202
"Dealer transactions, all other trading"	Includes transactions in which the firm acts as a principal in buying or selling debt (except US Treasuries) and derivative securities for the purpose of executing requested trades.	Corporate bonds, municipal bonds, agency bonds, mortgage-backed securities, asset-backed securities, collateralized mortgage obligations, commercial paper, certificates of deposit, Yankee bonds, foreign debt, options, warrants, futures, forwards, swaps	
"Investment banking services, advisory services"	Includes mergers and acquisitions and other advisory services. Advice and assistance are provided to firms that are merging, acquiring other firms, or being acquired, leveraged buyouts, corporate restructuring, and the reorganization of bankrupt and troubled companies.		5231103
"Investment banking services, underwriting services"	Includes all services related to the process of distributing new securities to investors, either through the public markets or to a private pool of investors.		
"Reverse repurchase agreement services"	Includes reverse repurchase agreement services, typically for a treasury note, agency bond, agency mortgage-backed security, or an investment grade corporate bond.		5231104
"Securities loan services and all other securities dealing services"	Includes securities lending services and all other securities dealing services.		

Transactions from the category for "Dealer transactions, equity securities" were used to calculate the index line *Dealer transactions, equity securities*. Those from the categories "Dealer transactions, Treasury securities and Dealer transactions, all other trading" were used to calculate the index line *Dealer transactions, debt securities and all other trading*. Similarly, the *Investment banking services* index line was composed of transactions from the categories for "Investment banking services, advisory services" and "Investment banking services, underwriting services." Transactions from the final two categories,

"Reverse repurchase agreement services" and "Securities loan services and all other securities dealing services" were included in the index line *Other securities dealing services*.

For each broad category selected through disaggregation, a BLS economist worked with the respondent to disaggregate further in order to select a unique transaction because the initial categories themselves are quite broad. For example, the "Dealer transactions, all other trading" category includes the dealing of all non-treasury debt and derivative securities. Therefore, additional disaggregation had to take place among corporate bonds, municipal bonds, agency bonds, etc. Once one of these types of securities was selected, disaggregation continued among the most heavily-traded specific securities within the selected type. The resulting unique transaction was a dealing transaction for a specific debt security. BLS would then ask the respondent to provide the price received for dealing this security each month. The price received by dealers is a bid-ask spread; that is, the difference between the price at which the dealer would sell a given security (the ask) and the price at which the dealer would buy it (the bid) if it were transacted on the current market.

2.3 Challenges with calculating index estimates

Using these traditional data collection procedures led to challenges calculating PPIs for the *Investment Banking and Securities Dealing* industry. Resource, sample size, and respondent burden constraints meant BLS collected prices for limited numbers of transactions from each establishment, putting a tremendous representativeness value on each item. For instance, the price movement for a single bond dealing transaction was meant to represent all price movement for the broader category of "Dealer transactions, all other trading" from that establishment. To give an idea of the magnitude of transactions this singular, specific securities dealing transaction would represent, the volume of corporate bonds traded in January 2019 alone was $32.4 billion and the total number of trades of municipal bonds exceeded 871,200 – and there are many other types of debt securities also included in this category.[6] BLS recognizes that for extremely high-transaction volume industries like *Investment banking and securities dealing*, using the typical PPI methodology of collecting the price each month for a single transaction does not provide the best estimate of price change for the millions of transactions that occur that month. Yet at the time other data collection options for this industry were not available for the PPI survey.

3. The search for alternative data sources

As price data for the *Investment banking and securities dealing* industry rolled in each month, BLS economists began to search for an alternative source of data that would provide a large number of accurate and representative transactions and prices while remaining within BLS' legal and budgetary constraints, creating no additional burden on respondents, and fitting into the monthly PPI production cycle. This last requirement, meeting the practical aspect of PPI production, involves the timeliness of processing data especially as it involves review and analysis by BLS economists. BLS explored multiple sources of alternative data for the industry, including regulatory organizations and private businesses. Immediately, the economists recognized any new data source would replace all or part of a broad service category described in Table 2 above rather than replacing the detailed index lines listed in Table 1, except where there was a one-to-one correlation. As time progressed, once they found and implemented an alternative data source, one of two paths followed. Either they would blend the alternative data with the existing directly-collected data provided by respondents or they would discontinue the directly-collected

[6] "U.S. Corporate Bond Trading Volume." *SIFMA*. https://www.sifma.org/resources/research/us-corporate-bond-trading-volume. Accessed 10 September 2019.

"U.S. Municipal Trading." *SIFMA*. https://www.sifma.org/resources/research/us-municipal-trading. Accessed 10 September 2019.

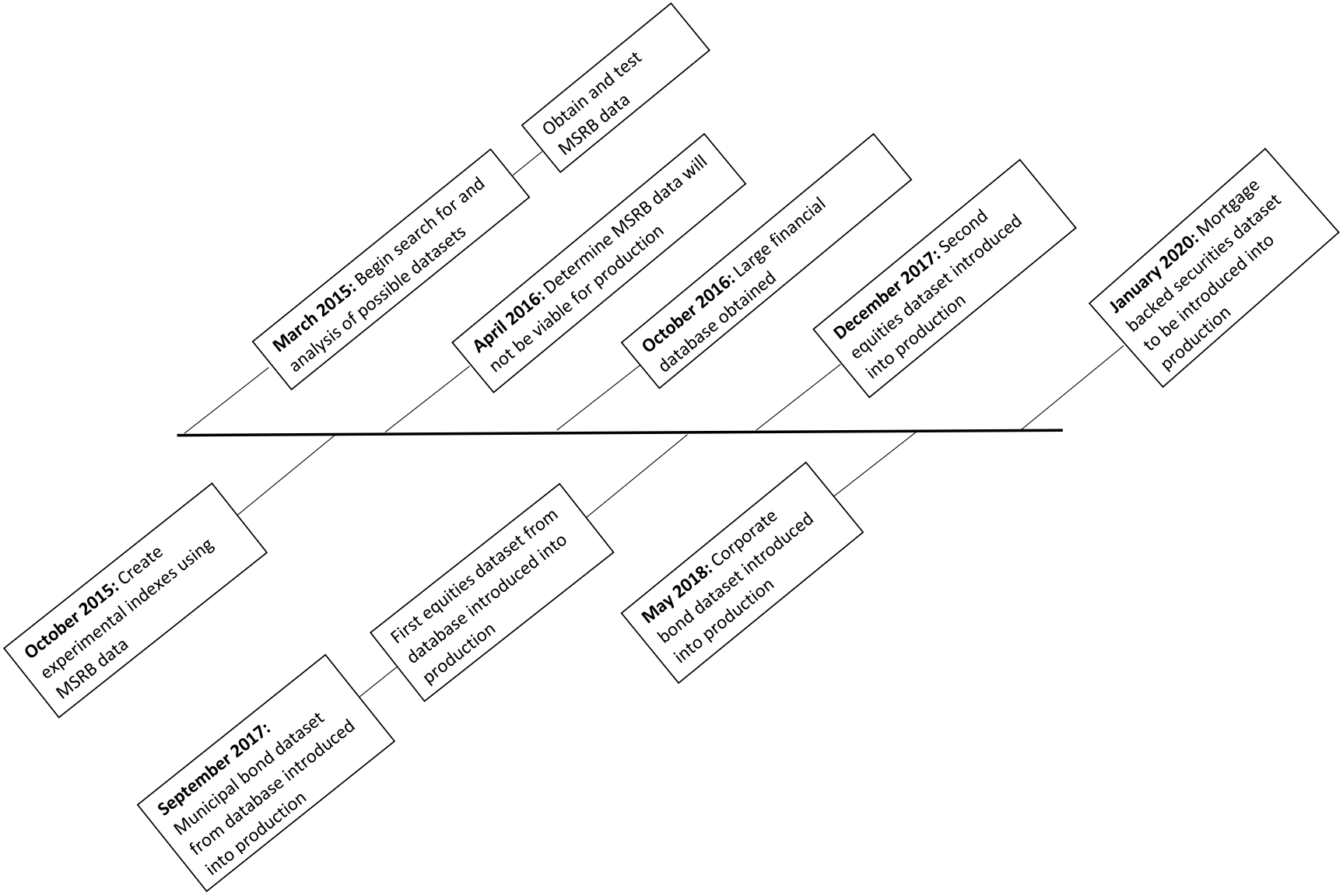
data and replace it with the alternative data. For the second path, they would reach out to the respondents and ask to discontinue monthly data collection in one category in order to increase it in others and/or increase cooperation. This requires further explanation.

The largest firms in this industry, and therefore ones from which the PPI survey needs cooperation most, conduct business in more than one of the industry's broad service categories. For instance, a firm may provide both equities dealing and investment banking services. If BLS could replace equities dealing with an alternative data source, then staff economists could offer to discontinue the monthly request for equities data in favor of more information on investment banking. Or, they could simply offer to discontinue the equities request, thereby reducing respondent burden, in a plea for continued cooperation in investment banking.

Finally, starting in March 2015, BLS intensified its efforts with big data, in part to be prepared for the fourth sample of this industry, which BLS should start collecting in late 2020 or early 2021. At that time, BLS will not need to address any broad service category or parts thereof for which the PPI survey has a viable alternative dataset. Staff economists will not request this data while initiating companies into the survey, which also reduces respondent burden and therefore hopefully will increase cooperation.

The next sections describe each case where BLS started using an alternative set of monthly price data for a broad service category. While some of these cases began in the early 2000s, Timeline 1 provides a perspective on the most recent efforts since work on different areas occurred concurrently.

TIMELINE 1:


Obtain and test MSRB data
March 2015: Begin search for and analysis of possible datasets
April 2016: Determine MSRB data will not be viable for production
October 2016: Large financial database obtained
December 2017: Second equities dataset introduced into production
January 2020: Mortgage backed securities dataset to be introduced into production
October 2015: Create experimental indexes using MSRB data
September 2017: Municipal bond dataset from database introduced into production
First equities dataset from database introduced into production
May 2018: Corporate bond dataset introduced into production

3.1 Process of finding an alternative data source – "Dealer transactions, equities securities"

In the early 2000s, BLS obtained two alternative data sources to replace directly-collected equities securities dealing transactions for the PPI. A respondent from a single source provided one of the datasets for free via e-mail. It listed over 2,000 ticker symbols along with corresponding closing bid prices, ask prices, and trade volumes for three days spread throughout each month (one day each from early in the month, the middle of the month, and late in the month in order to sample market movements throughout while also keeping data extraction burdens for the respondent low). BLS economists tested the data for use in PPI indexes and found that it addressed the methodological issues related to index accuracy of sample size and representativeness of the transactions. However, it had several operational risks related to revisions and response. If the respondent did not provide the data in a timely fashion each month, the PPI was at risk of not publishing the *Dealer transactions, equities* index because a large amount of the data needed to calculate price change would be missing. BLS would then need to revise the PPI index in a subsequent month once the data was received. If the respondent left the source organization and did not provide an alternative contact or simply decided to stop sending the data, the PPI might not be able to publish the index for months until a different data source was found or traditional data collection could be reinstituted. These scenarios could lead to a break in the index series.

BLS obtained the second alternative dataset from a website each month for a subscription fee. Similar to the first equities dataset, this one contained a different list of over 2,000 ticker symbols along with all of the same corresponding data for the same three days each month. BLS economists tested the data for use in PPI indexes and found it viable. However, there were challenges with purchasing this data set each year as its price changed in unpredictable ways and the source organization refused to register in the government procurement system, complicating the purchasing process. For instance, from 2015 to 2016, the price of the dataset more than doubled, making it very difficult to estimate a budget for the data. As with the first dataset, if BLS could no longer obtain this dataset due to rising costs and associated procurement regulations, index publishability would be at risk until a different data source was found or traditional data collection could be reinstituted. Still, BLS accepted the risks and moved forward in the early 2000s with using both data sources for equities as there were few options available providing such robust data, and in fact no serious problems materialized. However, because of these risks, BLS would continue looking elsewhere for equities data sources with at least similar robustness that would also be more stable.

3.2 Process of finding an alternative data source – Escalations for "Investment banking services, advisory services" and "Investment banking services, underwriting services"

For both of these broad service categories, the price BLS collects for the PPI is an estimated fee based on a percentage of the deal value, transaction value, or value of the offering for which the advisory or underwriting services are performed. Investment banks perform many services for many clients. According to the 2012 Census, investment banks earned over $13.9 billion in revenue from mergers and acquisition financial consulting services.[7] However, firms never provide the same exact service twice. Therefore, it can be difficult for them to estimate the value of the initially-sampled service on a monthly basis. In order to address this challenge, BLS economists researched alternative data that could be used to estimate the change in value of the sampled service rather than relying on the respondent to do so. They found that for the PPI they could use a combination of three sources: the Wilshire 5000, gross domestic

[7] "Finance and Insurance: Subject Series – Product Lines: Product Lines Statistics by Industry for the U.S.: 2012." *United States Census Bureau*, 5 February 2016, https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk.

product (GDP), and one of the Barclay's Bond Indexes to estimate the current day value. While the first two had no associated costs, in January 2007 BLS began purchasing Barclay's Bond Index data. Each month, once they escalated the deal value, transaction value, or value of offering using one or more of the three sources to better reflect the current period value of the initial pool of assets, BLS economists then blended this data with the percentage fee directly collected from sampled investment banks in order to calculate an estimate of price change. By asking only for the percentage fee each month, BLS reduced the burden on firms participating in the PPI survey.

3.3 Process of finding an alternative data source – "Dealer transactions, all other trading"

In order to find an alternative data source for debt securities dealing, in March 2015 BLS began exploring a number of different data sources including Markit, Bloomberg, MarketAxess, DelphX, IDC-Vantage, FINRA, and the Municipal Securities Rulemaking Board (MSRB). Staff economists evaluated various aspects of each data source including available data, accessibility, timeliness of data, format of data, and cost. After completing this analysis, they decided to further explore MSRB data for municipal debt securities.

BLS obtained sample datasets from the MSRB on a research basis, and by October 2015 staff was able to develop an initial matched-model index for municipal securities dealing using such characteristics as dated date, maturity date, interest rate, principal amount, source of repayment, rate type, and whether the security was callable, federal taxable, or had floating rate features. They presented these experimental indexes to staff at the MSRB and collaboratively discussed the type of data needed for the model, how it would be used on a production basis, and the confidentiality and security protections that would apply. However, in April 2016 MSRB determined it was unable to share the data required for production use in the PPI due to the terms under which MSRB is provided the data. Specifically, many of the descriptive data elements pertaining to the securities were provided under their CUSIP (Committee on Uniform Securities Identification Procedures) license and could not be shared on an ongoing basis. BLS investigated the cost of procuring a CUSIP license for the PPI, but the cost was prohibitive. Thus, BLS concluded it would have to continue looking elsewhere for debt securities data sources.

3.4 A solution – A single source for equities, some debt securities, and investment banking escalators

During their search in 2015-16, BLS economists identified a large alternative database for potential purchase that provides: daily closing bid-ask spreads for all securities traded throughout each month on U.S. exchanges; daily closing bid-ask spreads for municipal, corporate, and other debt securities; bond data for escalations; as well as other financial services data. BLS obtained this large database in October 2016. After vetting and experimental calculations, they determined that this single-source provided the data necessary to address the methodological concerns in the "Dealer transactions, all other trading" broad service category and met BLS' legal, budgetary, and operational requirements. Procuring regular access to this big database would make progress towards addressing the methodological concern of representativeness by replacing the directly-collected data for municipal bond dealing starting in September 2017 and for corporate bond dealing starting in May 2018. Thus, BLS could stop asking respondents for two of the most important types of products in the "Dealer transactions, all other trading" broad service category. The PPI survey would still need to directly-collect data from firms providing other products in this category, for instance mortgage-backed securities and options. With the introduction of data from this source, the PPI increased its municipal bond dealing data collection from one day per month to a full month and increased the number of observations per day by 6,640%. For corporate bond dealing, the PPI increased data collection from one day per month to a full month and

increased the number of observations per day by 3,971%.

This large database also resolved the operational issues with the previous two alternative data sources the PPI used for the "Dealer transactions, equities securities" broad service category, in addition to addressing methodological requirements. With the introduction of data from this source into production PPI indexes in two steps, the first in September 2017 and the second in December 2017, the PPI increased the number of days used in price calculation from three days per month to approximately 20 days (a full business month). The number of equities included in PPI samples was comparable between the alternative sources. BLS introduced data from the new large data source in two steps because 1) it placed a priority on replacing the PPI's purchased equities data source after its cost rose exponentially and 2) staff economists needed time to develop the data extraction software and modify existing code used to clean the data, analyze it, and calculate an estimate of price change.

Finally, BLS no longer needed to purchase the Barclay's Bond Indexes to estimate the current day value of sampled transactions in the two investment banking broad service categories because the large database provides sufficient bond data for escalation. This allowed BLS to put the funds allocated for Barclay's towards the cost of the new single-source database. Whether for escalations, equities, or debt securities, using this large database requires skilled staff to develop and maintain programs to categorize data into the proper broad service category and to process the data in accordance with established PPI methodology for this industry.

4. Incorporating this single-source alternative data into the Producer Price Index

Prior to implementation in production PPIs, BLS economists calculated volume-weighted average bid-ask spreads for each type of debt security for six months and compared them to the bid-ask spreads from the directly-collected data used in production indexes. The comparison was difficult because the price changes in the directly-collected price data for the individual debt securities were influenced by specific movements within related markets rather than overall market movement as a whole. However, results indicated that price movement calculated from the greater volume of transactions in the single-source alternative data seemed more representative of the industry movement as a whole due to the inclusion of thousands of additional data points representing a much larger part of the entire market.

To give a sense of the production processing involved after the PPI receives this large database each month, consider that, upon receipt BLS economists then extract the data into Excel spreadsheets and use a SAS program to remove data anomalies and calculate volume-weighted average bid-ask spreads for each type of debt security and equity exchange. They then manually enter these average spreads into the PPI's information technology system to be used in the calculation of the PPI's index estimates.

In terms of additional maintenance over time, note that staff created the initial lists of municipal and corporate bonds by selecting all bonds that met screening criteria for the PPI. Subsequently, they developed the screening criteria to ensure that the bonds selected are actively and consistently being traded and have valid pricing data in accordance with the matched-model concept. They also refresh the lists of bonds when the number of bonds for which there is data falls below 75% of the original number. Staff also removes bonds from the list as they reach their maturity date or if they are no longer being traded.

Prior to implementation in production PPIs, for each equity exchange BLS economists compared data from the new source to the data from the two original sources for six months to ensure that lists of equities obtained from each source was comparable and to evaluate the change in methodology of using a full month of data instead of the three days used previously. Results indicated that utilizing data from the

entire month was more accurate because the market can move quite differently on various days throughout the month, and now all of those movements are included in the overall price change calculation for the month.

BLS economists concluded that procuring and implementing this viable database results in improved PPI index estimates for the *Investment banking and securities dealing* industry because it now represents a larger proportion of the securities and equity dealing markets. An additional benefit is that greater representation occurs without having to rely on a large number of respondents to provide the data each month, thus reducing respondent burden. Procuring this database on an annual basis results in a net cost benefit to BLS, its procurement meets legal requirements, and its operational requirements are within the bounds that BLS can handle.

5. Retail trade

The importance of Retail trade industries and its sister sector, Wholesale trade, to the U.S. Producer Price Index could be the subject of another paper or two. This paper leaves aside the complexities of pricing these industries but provides a comparison of the BLS approach towards alternative data in the *Supermarkets and other grocery stores* retail trade industry with the approach used in the *Investment banking and securities dealing* industry. However, note that the PPI uses margin prices to estimate price change for Retail and Wholesale trade industries. Therefore, BLS requests detailed component data from survey respondents, including acquisition price and sale price, in order to determine the margin. The margin for a single transaction of a single product during the month sometimes fails to provide the best representativeness of price change for its product category. In some cases, even the average margin over the month for a single product may not provide the best representativeness for products in these extremely high-transaction volume industries that also have wide-ranging product categories.

Table 3 shows the U.S. PPI calculation structure for NAICS 445110, *Supermarkets and other grocery stores* retail trade industry:

TABLE 3:

Index Code	Index Title
445110	Supermarket and other grocery store services
445110P	Primary services
4451103	Supermarket and other grocery store services
44511032	Retailing of food and beverage products
445110321	Retailing of fresh meats
445110322	Retailing of fresh fruits and vegetables
445110323	Retailing of bakery products
445110324	Retailing of dairy products
445110325	Retailing of deli products
445110326	Retailing of frozen food products
445110327	Retailing of alcoholic beverages
445110328	Retailing of all other food and beverage products
44511033	Retailing of nonfood products
445110331	Retailing of cleaning and paper products
445110332	Retailing of health and beauty products
445110333	Retailing of all other nonfood products
445110SM	Other receipts

not all indexes are published

Looking at this industry in more detail, the index line *Retailing of fresh meats* has many of what the PPI survey terms "broad product categories" (the corollary of "broad service categories" in the *Investment banking and securities dealing* industry). For instance, the index may have beef, pork, poultry, fish, and seafood broad product categories. For data collection, a BLS economist would request that a respondent provide average margins for a broad product category such as all poultry transactions during the month. (Operationally this works out that the respondent provides the average acquisition price and the average sale price.) This broad product category average margin more accurately represents the index for *Retailing of fresh meats* than does one transaction of a pack of chicken wings during the month, or even the average margin for all transactions of the pack of chicken wings during the month. The next step of data collection becomes tricky. Through the standard PPI disaggregation process, if the next broad product category chosen for collection is in a different index line, perhaps *Retailing of fresh fruits and vegetables*, then for that company the price change estimate for *Retailing of fresh meats* is represented only by the price change for all poultry transactions. This is not ideal. A solution is for BLS to use big datasets provided directly by firms to address issues with both high-volume transactions and wide-ranging broad product categories and thus improve representativeness in the PPI *Supermarkets and other grocery stores* retail trade industry.

In early 2016 during the course of resampling the *Supermarkets and other grocery stores* industry, economists discussed big data with a few respondents who offered to send large datasets that they [more or less] had readily available. After vetting the data, economists recognized that the datasets contained all the necessary data elements – and they could ignore superfluous data – so this broader dataset replaced the narrower single product margin data that served as the fallback data collection option. There are two "wins" here: the establishments in question reduced their burden by providing a large dataset of records already on hand, and the representativeness of margin prices used in PPI estimates improved. Of course, the burden of calculating the margins shifted from the company to BLS staff, which must meet the operational requirements of processing the data within the PPI's production cycle timeframe. Using these large datasets requires skilled staff to develop and maintain programs to categorize data into the proper broad product categories and to process the data in accordance with established PPI methodology for this industry.

This experience was a success, although one that was not planned. In this case, BLS took advantage of accepting big data and incorporated it into the PPI's production process. One company now provides to BLS a dataset with acquisition and sales prices from which the PPI can compute the margin information for over 30,000 products covering 15 broad product categories. Another respondent provides a large dataset with transactions for the top 10 products for 11 broad product categories.

Beyond its happenstance nature, we include this brief look at the Retail trade case for two reasons. First, it further illustrates the challenges of measuring price change in industries with high-transaction volumes and explains the challenges posed by collecting data in wide-ranging product categories. There may be some lessons BLS can learn from the Retail trade experience that it can apply to Financial services. Second, this case demonstrates how BLS is utilizing an alternative data collection method rather than an alternative data source, an experience that might also apply to Financial services.

6. Legal and budget constraints[8]

We mentioned previously that the alternative datasets tried and/or implemented into production met BLS

[8] Directly from the working document *The BLS Framework for Alternative Sources and Collection Methods of Price Data*. September 13, 2019. To be published on the BLS website, www.bls.gov, in January 2020.

legal and budgetary requirements, or we explained where they failed. To set the context further, we note that just as with standard data collection, BLS follows procedures that comply with the Confidential Information Protection and Statistical Efficiency Act[9] (CIPSEA) for alternative sources and collection methods of price data. Accordingly, BLS pledges confidentiality, promising to use respondents' and third-party providers' data exclusively for statistical purposes. Until BLS secures permission from respondents, it cannot proceed with any type of data collection.

In the case of vendor-provided secondary source data, BLS often must negotiate contracts that are consistent with Federal laws (such as the number of option years BLS can have on a contract), that meet the needs of both parties, and that ensure costs are reasonably controllable in the longer term. Occasionally, a condition of the contract could be that the vendor be acknowledged publicly, and BLS can agree to this condition.

Although it does not apply to these case studies of PPI data, note that in the case of web/mobile data, Terms of Service (TOS) agreements for websites and Application Programming Interfaces (APIs) often have aspects that are problematic for Federal agencies. TOS often require acceptance of the establishment's state law over Federal law, and many TOS have open-ended indemnity clauses, two conditions to which Federal agencies cannot legally commit. As mentioned above, BLS provides the CIPSEA pledge to website owners and obtains consent to collect web/mobile data using BLS in-house software and/or the website's API with the understanding that the agency will use best practices and, if they have a TOS, explains which terms BLS will not be able to follow and why.[10]

As for budget requirements, in general BLS strives to assure that the transition from traditional, standard data collection to new alternative sources and collection methods of price data does not increase its overall budget, i.e. that this work remains at least budget neutral if not actually resulting in cost savings. There can be exceptions to this in situations where the gains to index accuracy, expanded coverage, and/or new products resulting from the use of alternative data sources or collection methods clearly outweighs any *net* increase in costs.

7. Conclusion – plans for the future

BLS must obtain big datasets in order to improve the representativeness of the Producer Price Index for industries with high-transaction volumes and /or wide-ranging product categories like *Investment banking and securities dealing* and those in the Retail trade sector. These may be new data sources, such as the large financial database BLS procures for *Investment banking and securities dealing*, or new data collection methods, such as obtaining large datasets from companies willing to participate in the PPI survey. As demonstrated with the original alternative data sources for equities, it may take several attempts to find the best data source. As demonstrated with data sources for securities dealing, first attempts may not be successful. A not trivial component is the human one: BLS must ensure its economists have the right data science skills in addition to economic prowess in order to optimize the alternative sources and methods that they find, especially within the operational requirements for creating economic statistics that are published monthly.

The successes outlined in this paper – and there are others – create a foundation which BLS can and will build upon for improving its PPI estimates. Among the lessons-learned from the foray into requesting and receiving large corporate datasets is that it is incorrect to assume that every company invests in its

[9] CIPSEA, 44 U.S.C. ch. 35, subch. I § 3501 et seq.

[10] At the request of the respondent, BLS is prepared to document its approach to the company's TOS and formalize a written agreement.

information technology infrastructure such that it has “easy” access to the kinds of data that the PPI requests. There are more approaches to alternative data for BLS to try, and there are methodological questions that still need answers.

For instance, there are many products in the “Dealer transactions, all other trading” broad service category. The PPI is now measuring monthly price change for corporate and municipal bonds from data in the large database it purchases. It still directly collects data for the other products, such as mortgage-backed securities, certificates of deposit, and options. Is this necessary? Or, are the markets for these various products tied closely enough to corporate and municipal bonds that BLS can assume that price change trends similarly? BLS economists will be examining this potential simplification. Similarly, BLS is searching for an alternative data source for the “Dealer transactions, Treasury securities” broad service category, which seems promising.

For Retail trade industries, BLS will ask companies that are already providing prices monthly if they can provide datasets with more transactions for the PPI. In this case, BLS cannot simply calculate margins and add them to the PPI, so it is actively considering how best to weight transactions that enter the survey through alternative datasets after the industry sample is established. BLS already has real world PPI examples to evaluate.

Finally, in spring 2020 BLS will start collecting new PPI samples for two industry groups in the Wholesale trade sector, NAICS 4244, *Grocery and Related Product Merchant Wholesalers* and NAICS 4248, *Beer, Wine, and Distilled Alcoholic Beverages Merchant Wholesalers*. BLS will ask some companies to provide big datasets of transactions across broad product categories rather than asking for a handful of representative average product transactions in a limited number of product categories as decided by the disaggregation process. This will undoubtedly provide new and useful experiences.

References

The BLS Framework for Alternative Sources and Collection Methods of Price Data. September 13, 2019. To be published on the BLS website, www.bls.gov, in January 2020.

2011. "Chapter 14. Producer Prices" in *BLS Handbook of Methods*. https://www.bls.gov/opub/hom/pdf/ppi-20111028.pdf

2019. "Quality Adjustment in the Producer Price Index" on the BLS website, *bls.gov*. https://www.bls.gov/ppi/qualityadjustment.pdf

Confidential Information Protection and Statistics Efficiency Act of 2002 (CIPSEA). United States Code. Title 44, chapter 35, subchapter I, section 3501 et seq. https://www.bls.gov/bls/cipsea.pdf.

"Finance and Insurance: Subject Series – Product Lines: Product Lines Statistics by Industry for the U.S.: 2012." *United States Census Bureau*, 5 February 2016, https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk.

"U.S. Corporate Bond Trading Volume." *SIFMA*. https://www.sifma.org/resources/research/us-corporate-bond-trading-volume. Accessed 10 September 2019.

"U.S. Municipal Trading." *SIFMA*. https://www.sifma.org/resources/research/us-municipal-trading. Accessed 10 September 2019.