

The 34th Meeting of the Voorburg Group on Service
Statistics

Paris, France, 30 September – 4 October 2019

Poster Session by Bonnie Murphy, U.S. Bureau of Labor Statistics

Implementing an Alternative Data Source to Estimate Producer
Price Indexes within Selected Financial Services Industries

Melanie Santiago

U.S. Bureau of Labor Statistics



Introduction (background from the working paper “BLS Framework for Alternative Sources and Collection Methods of Price Data”)

The U.S. Bureau of Labor Statistics (BLS) generally relies on its own sample surveys to collect the price and weighting information necessary to estimate the Producer Price Index (PPI). For decades, BLS has utilized alternative sources of price data as well, but on a limited basis. Now, the burgeoning availability of “big data” compiled by other government agencies and private businesses provides the *potential* to measure price change more accurately, improve BLS’ management of respondent burden, expand item and geographic coverage, publish new products, and/or achieve cost savings.

Calculating price indexes based on traditional data collection presents several methodological challenges that might be further mitigated or resolved through the use of new alternative data sources and collection methods. First, because BLS price indexes measure constant quality price change over time, when a unique item is no longer available a replacement item may need to be selected, and any quality change between the original and replacement item must be estimated and removed to reflect pure price change. Second, new products entering the marketplace must be accounted for in a timely manner with the appropriate weight. Third, BLS price indexes are based on samples, which are subject to sampling error. Selection of a unique item using a multistep probability selection technique is ideally made with information from the respondent. When access to the respondent is limited, BLS uses “fallback” procedures to estimate proportions for selecting a unique item, but the procedures create a less representative sample than desired. Fourth, BLS price surveys have minimum quality criteria for publishing index estimates. At a minimum, publishability depends on the number of establishments participating in the survey and the number of items for which they provide prices, which must be sufficient to ensure representativeness and protect against respondent identification. Other criteria, including standard errors, could factor into this decision as well.

As well, BLS faces challenges related to survey operations. Response rates are declining due to many factors such as: confidentiality requirements, increasing number of competing surveys, and data security concerns. Data collected by BLS through monthly pricing surveys is increasingly costly. It takes more time and perseverance to contact potential survey respondents and secure cooperation. Diminishing data collection resources result in fewer establishments and items in the survey, which negatively impacts index quality.

Adopting new alternative sources and collection methods of price data may allow BLS to mitigate or resolve some of these methodological and operational challenges. The observations provided by alternative data sources and collection methods often far outnumber those from traditional data collection. As long as the alternative data source is comparably representative of the target population, a larger sample size may allow BLS to measure price change more accurately through decreased sampling error, by better reflecting substitution patterns including the timely introduction of new items, and by better accounting for quality change. BLS is adapting to this new alternative data environment, recognizing 1) that this may modify the

“matched model” concept for price measurement by employing it only at a higher level of aggregation in selected product categories and/or industries, and 2) that it *could* result in skipping steps of the probability selection technique, which already occurs with current disaggregation fallback methods. Furthermore, alternative data may help BLS price surveys reduce and/or better manage respondent burden, address survey non-response, reduce collection costs, and publish more indexes or publish at a more detailed level. More generally, we need to be mindful that use of alternative data and/or methods present their own set of significant challenges and thus require careful evaluation before decisions are made on using them in production.

Alternative data collection also presents and introduces its own set of methodological and operational challenges, many of which stem from the fact that BLS does not have *control* over the data. BLS must ensure that data from any alternative source meets:

- PPI’s core measurement objectives and is representative of its target population;
- BLS requirements for *data continuity* (a sudden loss of an alternative data source cannot cause an index to drop out of publication, and the data elements and structure of alternative datasets cannot cause sudden and urgent rework to BLS’ Information Technology (IT)); and
- The agency’s ability to be transparent, including ensuring *index reproducibility* (see OMB’s *Information Quality Act*, Public Law 106-554 § 515).

Budget Requirements

In general, BLS strives to assure that the transition from traditional 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.

Legal Constraints

Just as with traditional data collection, BLS follows procedures that comply with the *Confidential Information Protection and Statistical Efficiency Act* (CIPSEA, 44 U.S.C. ch. 35, subch. I § 3501 et seq.) for all alternative sources 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.

In the case of web scraping data, Terms of Service (TOS) agreements for websites and 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 web scrape 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.

BLS has undertaken several projects in an attempt to supplement and/or replace its traditional direct collection of price data with new alternative sources. One area in which BLS has been particularly successful is in the financial services sector. PPI is now using a large, purchased dataset to estimate PPIs for financial services within the U.S. economy. PPI extracts thousands of data points per day to compute a weighted average price for use in index estimation. PPI is using this source to replace directly-collected data for corporate securities dealing, municipal securities dealing, and equities securities dealing in the investment banking industry. In addition, it is blended with directly-collected data to escalate values in the investment banking and in the investment advice industries.

Investment Banking and Securities Dealing Industry

Industry Definition

Investment banking and securities dealing (NAICS 523110), as defined by the 2012 North American Industry Classification System, are those establishments which are primarily engaged in underwriting, originating, and/or maintaining markets for issues of securities. Investment bankers act as 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 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, 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:

- Establishments primarily engaged in acting as agents (i.e., brokers) in buying or selling securities on a commission or transaction fee basis are classified in Industry 523120, Securities Brokerage.
- Investment clubs or individual investors primarily engaged in buying or selling financial contracts (e.g., securities) on their own account 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 turnover 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.

Product Structure

The following table shows the U.S. PPI structure for NAICS 523110, Investment banking and securities dealing.

Index Codes	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

The U.S. PPI for investment banking and securities dealing closely follows the NAPCS structure, although the NAPCS structure includes more detail in several areas.

The service line “Dealer transactions, equities” includes:

- All dealer spreads earned on equities trades, including those that occur on an exchange and those that occur in the over-the-counter (OTC) market.

The service line “Dealer transactions, debt securities and all other trading” includes:

- All dealer spreads earned on trades of corporate, treasury, and municipal debt securities, options and other derivative securities, and all other non-equity securities.

The service line “Investment banking services” includes:

- Underwriting of new issue securities to be placed in public markets
- Underwriting of new issue securities to be placed in private markets
- Securitization of assets. This is the issuance of mortgage-backed securities (MBS), asset-backed securities (ABS), and collateralized debt obligations (CDOs).
- Mergers and acquisitions (M&A) and other advisory services. M&A advisory services include leveraged buyouts, restructuring and recapitalization of companies, and the reorganization of bankrupt and troubled companies.

The service line “Other investment banking and securities dealing services” includes:

- Stock loans (securities lending) - lending transactions in which securities are used as collateral. Securities loans in exchange for cash collateral are not eligible for collection since these transactions are not considered to be output generating.
- Reverse repurchase agreements (reverse repos) - when a sampled firm acts as the lender of funds (reverse repos), these transactions are collected with prices based on the interest income earned. Conversely, repos are not collected since they do not constitute turnover generating activity.

Measurement Issues

Several challenges arise in calculating PPIs in this industry using traditional data collection. Calculating more detailed service lines for securities dealing is challenging because there are many different types of securities and extremely large volumes of trades. Due to resource and respondent burden constraints, prices for limited numbers of transactions are collected from each firm. In order to have enough data to calculate a quality price index, various types of securities with similar price movements must be aggregated into broader index lines.

For example, the PPI for Dealer transactions, equity securities, tracks prices received by securities dealers for executing trades of equities. The price received by dealers is a bid-ask spread – 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 to be transacted on the current market. Using traditional PPI data collection methods, each sampled securities dealer would be asked to provide a bid-ask spread each month for the trade of a specific constant equity. Because the dealer would also be asked to provide spreads for trades of other securities as well, it would most likely be too burdensome to request more than one or two equities trades from each sampled dealer. For debt securities and all other trading, it would be highly burdensome to ask a firm to provide pricing data for three different types of bonds, several types of options and each type of derivative security they transact. Instead, each firm is asked to provide pricing data on a few transactions within the broader category of debt securities and all other trading.

Unfortunately, tracking the bid-ask spread for a few securities per dealer does not give us an

accurate estimate of price change for the millions of trades of all securities that take place each month. Therefore, the quality of our data is limited by sample size and reporter burden.

Selecting a single transaction that represents all transactions of a certain type can also be challenging. Often, a single municipal bond dealing transaction is not very representative of all municipal bond transactions completed by the securities dealing firm. At the industry level, the sample allocation and data collection resources are not available to collect a large enough sample to represent the millions of transactions that are conducted nationally.

Alternative Sources of Financial Services Data

In order to improve our ability to reflect price change for the Investment banking and securities dealing industry overall, an alternative financial data set that provides large quantities of equity and debt security transactions is being used in place of traditional PPI data collection. Incorporating large quantities of transactions into price calculations creates more representative price movements, while at the same time reducing respondent burden. BLS explored multiple sources of alternative data for the Investment banking and securities dealing industry, including regulatory organizations and private businesses.

Dealer transactions, equities

PPI initially used two alternative data sources to replace directly collected equities securities dealing transactions. One of the datasets was provided to us for free via email by a single representative. This method of data collection addressed the methodological issues related to index accuracy of sample size and representativeness of the transactions, but it had several operational risks related to revisions and response. If that representative did not provide us with the data in a timely fashion each month, the index was at risk of not publishing because a large amount of the data needed to calculate price change would be missing. A revision would then need to be done once the data was received. If the representative left the organization and did not provide an alternative contact or if he simply decided to stop sending the data, the index might not publish until a different data source was found or traditional data collection could be instituted.

The second alternative dataset was obtained from a website each month for a subscription fee. There were challenges with purchasing this data set each year, as the price changed in unpredictable ways and the organization providing the data refused to register in the government procurement system, creating problems for procurement staff. From 2015 to 2016, the price of the dataset more than doubled making it very difficult to estimate a budget for the data. Also, government policy does not allow purchases over the Government Purchase Card limit if the organization is not registered in the System for Award Management (SAM). As with the first dataset, if this dataset could no longer be obtained due to the rising costs and associated procurement regulations, index publishability would be at risk until a different source of data could be found or traditional data collection could be instituted.

Dealer transactions, debt securities

In addition to the operational risks associated with the two sources of alternative data for equities security dealing mentioned above, the methodological issues of sample size and representativeness persisted in the debt securities dealing index. PPI explored obtaining financial data from regulatory organizations such as the Municipal Securities Rulemaking Board (MSRB), which had datasets of actual transactions. After extensive discussions regarding the type of data needed, how it would be used, and the confidentiality and security protections that would apply, MSRB determined that they were unable to share the data due to the terms under which the data is provided to them.

Benefits of a single source for both equities and debt securities

PPI identified another alternative data source that provides daily closing bid-ask spreads for all securities traded during each month on U.S. exchanges, closing bid-ask spreads for municipal, corporate, and other debt securities, as well as other financial services data. This single source provides the data necessary to address the methodological concerns in the debt securities dealing index by replacing the directly collected data for corporate and municipal bond dealing and also resolves the operational issues with the previous alternative data sources used for the equities securities dealing index. With the introduction of data from this alternative financial data set, PPI increased the number of days used in price calculation for equities dealing from three days per month to approximately 20 days (a full month). The number of equities included in PPI samples was comparable between the alternative sources. For corporate bond dealing, PPI increased from one day per month of directly-collected pricing data to a full month. At the same time, PPI increased the number of observations per day by 3,971%. For municipal bond dealing, PPI also increased from one day per month of directly-collected pricing data to a full month and increased the number of observations per day by 6,640%. Use of the alternative data leads to a better quality index because a larger proportion of the securities dealing market is represented without having to rely on a large number of respondents to provide the data each month.

Steps to incorporate data into the price index

The initial lists of corporate and municipal bonds were created by selecting all bonds which met our screening criteria. The screening criteria was developed to ensure that the bonds selected were actively and consistently being traded and had valid pricing data. The lists of bonds are refreshed when the number of bonds for which there is data falls below 75% of the original number. Bonds are removed from the list as they reach their maturity date or if they are no longer being traded. Volume weighted average bid-ask spreads were calculated for each type of debt security and experimental indexes were calculated for nine months and compared with the indexes calculated using directly collected data prior to introducing the secondary source data into production.

For each equity exchange, data from the new source was compared with the data from the two original sources for 6 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.

Each month, the data is extracted into Excel spreadsheets and the industry analyst uses a SAS program to systematically remove data anomalies and calculate volume weighted average bid-ask spreads for each equity exchange and type of debt security. These average spreads are then manually entered into the PPI's repricing and estimation system at the item level to be used in index calculation.

Remaining challenges

While the use of alternative data has allowed us to better represent the securities dealing market as a whole and calculate more accurate indexes, some methodological challenges remain. Sampled transactions in the PPI are weighted by a measure of their size and importance. In the first stage of PPI computation, price indexes are constructed for narrowly-defined groupings of goods or services. The individual transactions included in these indexes are weighted by the producing establishment's turnover for the product line. In the second stage of PPI computation, indexes for individual goods and services are combined into aggregate indexes. Data for weighting together the product line indexes comes primarily from the Economic Censuses of the U.S. Census Bureau. These weights are updated every 5 years.

While weighting transactions based on establishment turnover from respondents and weighting indexes based on product line turnover from the U.S. Census is rather straight forward, weighting third party data sources creates a challenge. BLS researched total volume and volume weighted average spreads for an entire year for each equity exchange and type of debt security, and continues to research possible sources that can be used to weight third party data against the directly collected data it is blended with in various index lines.

Conclusion

The index quality improvement that comes from using alternative data in the Investment banking and securities dealing industry vastly outweighs the challenges involved. BLS will continue to look for ways to supplement and/or replace traditional data collection with data from new alternative sources and to use new alternative collection methods. BLS is establishing a framework that presents the case for considering alternative sources and collection methods and addresses common nomenclature, legalities, budgetary focus, and the need for vetting each data source (some of which is presented in the introduction). To ensure each alternative data set meets our core measurement objectives, BLS is establishing processes for evaluating the data for fitness for use and any trade-offs necessary to use the data including changes to our strategies for gaining respondent cooperation, to review procedures, and to IT applications. More generally, BLS is evaluating all impacts on business processes and developing a more standardized approach to handling alternative data.