Big Data in the U.S. Consumer Price Index: Experiences & Plans

by

Crystal G. Konny, Brendan K. Williams, and David M. Friedman

February 2019

Abstract

The Bureau of Labor Statistics (BLS) has generally relied on its own sample surveys to collect the price and expenditure information necessary to produce the Consumer Price Index (CPI). The burgeoning availability of big data has created a proliferation of information that could lead to methodological improvements and cost savings in the CPI. The BLS has undertaken several pilot projects in an attempt to supplement and/or replace its traditional field collection of price data with alternative sources. In addition to cost reductions, these projects have demonstrated the potential to expand sample size, reduce respondent burden, obtain transaction prices more consistently, and improve price index estimation by incorporating real-time expenditure information—a foundational component of price index theory that has not been practical until now. In CPI, we use the term alternative data to refer to any data not collected through traditional field collection procedures by CPI staff, including third party datasets, corporate data, and data collected through web scraping or retailer API's. We review how the CPI program is adapting to work with alternative data, followed by discussion of the three main sources of alternative data under consideration by the CPI with a description of research and other steps taken to date for each source. We conclude with some words about future plans.

Crystal G. Konny is Chief of Branch of Consumer Prices at the Bureau of Labor Statistics (BLS). Brendan K. Williams is a Senior Economist in the Branch of Consumer Prices. David M. Friedman is the BLS Associate Commissioner for Prices and Living Conditions. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the view of the U.S. Bureau of Labor Statistics. We thank Kate Sosnowski, Kelley Khatchadourian, Jason Ford, Lyuba Rozental, Mark Bowman, Craig Brown, Nicole Shepler, Malinda Harrell, John Bieler, Dan Wang, Brian Parker, Sarah Niedergall, Jenny FitzGerald, Paul Liegey, Phillip Board, Rob Cage, Ursula Oliver, Mindy McAllister, Bob Eddy, Karen Ransom, and Steve Paben for their contributions.

Big Data in the U.S. Consumer Price Index: Experiences & Plans

Introduction

The Bureau of Labor Statistics (BLS) has generally relied on its own sample surveys to collect the price and expenditure information necessary to produce the Consumer Price Index (CPI). The burgeoning availability of big data has created a proliferation of information that could lead to methodological improvements and cost savings in the CPI. The BLS has undertaken several pilot projects in an attempt to supplement and/or replace its traditional field collection of price data with alternative sources. In addition to cost reductions, these projects have demonstrated the potential to expand sample size, reduce respondent burden, obtain transaction prices more consistently, and improve price index estimation by incorporating real-time expenditure information—a foundational component of price index theory that has not been practical until now.

Government and business compile big data for their administrative and operational needs, and some of these data sources can be used as alternatives to BLS's surveyed data with some necessary adjustments. We use the term alternative data to refer to any data not collected through traditional field collection procedures by CPI staff, including third party datasets, corporate data, and data collected through web scraping or retailer API's.[1] Alternative data sources are not entirely new for the CPI. Starting as far back as the 1980's, CPI used secondary source data for sample frames, sample comparisons, and supplementing collected data to support hedonic modeling and sampling. What is new now is the variety and volume of the data sources as well as the availability of real-time expenditures. This chapter will review BLS efforts to replace elements of its traditional survey with alternative data sources and discuss plans to replace and/or augment a substantial portion of CPI's data collection over the next few years. After a brief overview of the CPI, we will review how the CPI program is adapting to work with alternative data, followed by discussion of the three main sources of alternative data under consideration by the CPI with a description of research and other steps taken to date for each source. The paper concludes with some words about future plans.

Overview of CPI

The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of goods and services. The CPI is a complex measure that combines economic theory with sampling and other statistical techniques and uses data from several surveys to produce a timely measure of average price change for the consumption sector of the American economy. BLS operates within a cost-of-living-index (COLI) framework when producing the CPI.

[1] Typically, "big data" is described as very large data sets including structured, semi-structured, and/or unstructured data that has the potential to be mined for information and subject to advanced analytical applications including machine learning. While some of the data sets the CPI program has been investigating meet this description, some are not that large but contain many of the same benefits and challenges presented by other private sector alternatives to survey data and thus we tend to focus more on the term "alternative data" rather than "big data" as described in the remainder of this paper. CPI is dealing with many of the potential benefits and challenges presented by the new business paradigms for Federal statistics agencies described in the two CNSTAT reports released in 2017 and 2018: https://www.nap.edu/resource/24893/Multiple%20Datasources.pdf .

Weights used in the estimation of the CPI are derived primarily from two surveys. The Consumer Expenditure (CE) Survey furnishes data on item category purchases of households and is used to draw the CPI item sample. The Telephone Point of Purchase Survey (TPOPS) collects data on retail outlets where households purchased commodities and services and is used as the outlet frame from which BLS selects a sample of outlets.[2] Weights are derived from the ratio of the probabilities of selection. BLS has not had access to the expenditure information necessary to produce superlative indexes, the preferred class of index formulas for COLI estimation, for the lower level component indexes that feed all CPI outputs. We currently only use a superlative index formula to produce the Chained CPI-U at the upper level of aggregation.[3] The lower level indexes used in CPI aggregates[4] almost all use a geometric mean index formula, which approximates a COLI under the restrictive assumption of Cobb-Douglas utility.

Pricing information in the current CPI is primarily based on two surveys. The Commodities and Services (C&S) survey is conducted by BLS data collectors, known as Economic Assistants (EAs), who visit each store location or website (known as an outlet in BLS nomenclature) selected for sampling. For each item category, known as an Entry Level Item (ELI), assigned to an outlet for price collection, an EA, using information from a respondent on the portion of the outlet's sales of specific items, employs a multistage probability selection technique to select a unique item from among all the items the outlet sells that fall within the ELI definition. The price of that unique item is followed over time until the item is no longer available or that price observation is rotated out of the sample. The Housing survey is used to collect rents for the Rent of Primary Residence (Rent) index and these rent data are also used to calculate changes in the rental value of owned homes for the Owners' Equivalent Rent index. While the CPI has generally used these surveys for price and rent data, historically in several cases we turned to alternative data sources, including for used cars and airline fare pricing and sales tax information.[5]

Several challenges arise in calculating the CPI using traditional data collection; we summarize the main ones that are relevant to our thinking about the use of new alternative data sources. First, because the CPI measures constant quality price change over time, when a unique item is no longer sold a replacement item must be selected, and any quality change between the original and replacement items must be estimated and removed to reflect pure price change in the index. Second, new goods entering the

[2] BLS is currently pursuing an effort to include the collection of point-of-purchase information within the Consumer Expenditure Survey. This will replace TPOPS starting with indexes released in FY 2021.

[3] See "Improving initial estimates of the Chained Consumer Price Index" in the February 2018 issue of the *Monthly Labor Review* for more information on changes made to the formula in calculating the *preliminary* C-CPI-U starting with the release of January 2015 data.

[4] The three main CPI series are the CPI for All Urban Consumers (CPI-U), the CPI for Urban Wage Earner and Clerical Workers (CPI-W), and the Chained CPI for All Urban Consumers (C-CPI-U). The CPI-U and its component indexes are the "headline" statistics in the monthly CPI News Releases. The same set of prices are used for all three families of indexes. The CPI-U and CPI-W use the same formula but different consumer expenditure weights to aggregate basic indexes while the C-CPI-U uses its own formula and weights. Major operational decisions in the CPI are made and assessed based on optimizing the quality of the CPI-U. For example, procedures to select the geographic sample, create samples of outlets and prices, and general decisions about resource allocation are made and analyzed with the CPI-U in mind, rather than, for example, the CPI-W or average price data. For additional detail on the construction of the CPI, see "Chapter 17. The Consumer Price Index" in *The BLS Handbook of Methods* (https://www.bls.gov/opub/hom/pdf/homch17.pdf).

[5] See the CPI Fact Sheet on "Measuring Price Change in the CPI: Used Cars and Trucks" for more information on the use of National Automobile Dealers Association (NADA) data for used cars beginning in 1987. Airline fare pricing was previously based on prices collected from the SABRE reservation system and is now collected using web-based pricing. See the CPI Fact Sheet on "How BLS Measures Price Change for Airline Fares in the Consumer Price Index" for more information.

marketplace must be accounted for in a timely manner with the appropriate weight. Third, the CPI is based on samples, which can introduce sampling error. Selection of a unique item using a multistage probability selection technique is ideally made with information from the respondent. When access to the respondent is limited, BLS uses procedures to estimate proportions for selecting a unique item, but the procedures may not create as representative a sample as desirable. In addition, the CPI may only be able to collect offer prices that might not reflect all of the discounts applied to a transaction. In terms of survey operations, data collected by BLS through pricing surveys is increasingly costly and more difficult to collect. Metropolitan areas have generally increased in size, which causes a corresponding increase in travel costs. The increase in the number of chain stores has increased the time to obtain corporate approval to collect data in stores. Response rates are declining due to many factors: new confidentiality requirements, increasing number of surveys, increasing distrust of government, data security concerns, and/or less confidence in the accuracy of the CPI.

Working with Alternative Data in CPI

Today's alternative data sources provide an opportunity to address many of the challenges encountered by the CPI over the past few decades. Adopting alternative data sources may allow us to more accurately measure price change. Alternative datasets may allow us to use larger sample sizes, transaction rather than offer prices, reflect consumer substitution patterns, remove quality change, reduce or eliminate respondent burden, help address non-response problems in the CPI's surveys, and reduce collection costs in some situations. In some instances, we are able to get real-time expenditure information as well. Data may be at a more granular level, for many more items than in our sample, or timelier such as daily. Initial exploration of the use of alternative data in CPI was focused on response problems and improving index accuracy in hard-to-measure product areas. In more recent years, BLS is giving equal attention to finding new cost efficiencies in the collection process.

The CPI program classifies its alternative data sources into three main categories:

1. *Corporate supplied data* are survey respondent provided datasets obtained directly from corporate headquarters in lieu of CPI data collectors collecting data in respondent stores. As the datasets are typically created for their own use, data elements and structure are defined by the respondent, and the BLS has to adapt them to our systems. BLS receives varying levels of information about the datasets – in general, the information provided is what the companies are willing to give us. Discussions with corporate data respondents often involve finding a level of aggregation that the corporation is comfortable providing to address confidentiality concerns.
2. *Secondary source data* (third party datasets) are compiled by a third party, contain prices for goods or services from multiple establishments, and need to be purchased by BLS or, in some limited cases, are provided free of charge from the data aggregator. The data aggregator has made some effort to standardize the data elements and structure across business establishments.
3. *Web scraping data* are data collected by BLS staff automatically using software that simulates human web surfing to collect prices and product characteristics from websites. Some establishments provide Application Programming Interfaces, or *APIs*, to allow partners to access pricing information. Data collection through an API is often easier and more straight-forward than maintaining web scraping code over time. In both cases, BLS follows procedures that are in compliance with the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).

BLS needs to evaluate each alternative data source regardless of type to ensure it meets the measurement objectives of the CPI as well as various operational considerations. Our experience with each of these data categories is described later in the Experiences section. In general, our process for deciding whether an alternative data source is fit for CPI usage currently involves the following steps for each item or establishment:

1. Determine what item or establishment to pursue (criteria we currently consider are reflected in the Appendix table)
2. Evaluate alternative data source options
3. Evaluate selected data source, including definition, coverage, and other quality aspects
4. Evaluate data quality over a predefined amount of time, which will depend on the type of data
5. Determine research approach and alternative methodologies to test, including
    - match and replace individual prices in CPI with individual prices in alternative source (see Wireless Telephone Services case)
    - match and replace individual prices in CPI with an average price for a unique item or over a defined set of items (see CorpY case)
    - replace price relatives in the CPI with estimates of price change based on new methodologies (see the CorpX and New Vehicles cases)
    - use all establishments and items in alternative data and calculate an unweighted index (see Crowd-sourced Motor Fuels case)
6. Evaluate replacement indexes based on statistical tests and cost benefit analysis
    - Our criteria for using in production: Is the data a good fit for CPI? Is it as good as or better than our current pricing methodology? Is it more cost effective or does the improvement in the index justify the additional cost? In some cases, BLS will implement a short-term solution that meets the criteria for use in production while still researching longer-term improvements (see Corp X case for example).
7. Determine the best way to incorporate the data into the CPI (e.g. transition plans, risk mitigation/contingency plans, systems considerations, etc.)

While we see numerous potential benefits in introducing new alternative data sources in the CPI as noted at the start of this section, we have also encountered challenges that have impeded the CPI from quickly incorporating alternative data into the program's outputs. Prior to discussing specific experiences, we will summarize the challenges/issues we have attempted to address in our pilot projects to date. Even uses that seem relatively straightforward, such as replacing manual collection of online prices with web scraping, can be complicated by legal and/or policy questions. Using alternative data often requires adopting new methodologies since index calculation methods are directly tied to BLS's survey and sampling systems and the conceptual foundation of the CPI program. Challenges faced by CPI in the use of alternative data can be grouped into three categories: those related to methodological needs of the CPI; those related to dealing with survey operations; and those related to legal and/or policy requirements.

Methodological Challenges

Because the CPI is designed to use the BLS's own surveyed data, we have encountered some challenges related to alternative data congruence with CPI methodology. In some cases, we have had to introduce new methods into the CPI to make full use of an alternative data source (see New Vehicles); in other cases, we had to relax some of our requirements (see CorpY); and in still others, we have had to decide that a particular source does not meet CPI measurement objectives (see Housing).

The primary obstacle to dealing with transaction data in the CPI has been dealing with *product lifecycle effects*, i.e. when products exhibit systematic price trends in their lifecycle. For certain goods such as apparel and new vehicles, a product is typically introduced at a high price on the market and is gradually discounted over time. At the point where the good exits, the price has been discounted substantially and may be on clearance. In the CPI, a similar good is selected, and its price is compared with that of the exiting good. The price relative constructed by comparing these two items typically implies a large increase in price from the exiting good to its replacement. In one large jump, this increase will offset the incremental price declines over the prior product's lifecycle. While this method works in the CPI's fixed weight index, a price comparison between exiting and new goods in a dynamically weighted index may under-correct in situations where an exiting item is a low inventory item on clearance or over-correct in other situations.[6] We found that multilateral price index methods designed to address chain drift[7] did not remedy downward drift associated with product lifecycles. Conventional hedonic methods also do not address product lifecycle effects.[8] We have found that hedonic price indexes often exhibit the same drift as matched-model indexes (see Greenlees and McClelland). For new vehicles, we developed a method of using year-over-year price change to avoid lifecycle effects. More generally, the implications of product lifecycles have not received much attention in the price index literature with some exceptions such as Melser and Syed (2015).

Many alternative data sources are constructed as "convenience" samples, based on the ease of collecting data on a certain segment of the market. When major companies, brands, or market segments are not represented in an alternative dataset, it can suffer from *loss in representativeness*, thus potentially introducing coverage error into the CPI that is based on representative samples. Comingling sampled and unsampled data can undermine the interpretation of the CPI's existing variance measurement, which in addition to providing a measure of the uncertainty in the CPI due to sampling, is used to allocate the CPI's sample across items and outlets to minimize variance.[9] An inaccurate estimate of variance could cause an inefficient allocation of the sample. This is less of an issue in cases where we are totally replacing survey collection with an alternative data source such as in the New Vehicles case. In other instances, such as with scanner data, where the source can be seen as actually representing the universe of products sold and thus a better reflection of the real world, CPI can treat the data as if it were produced by stratified random sampling or adapt variance estimation to reflect no contribution to sampling variance from the given component. While big data presents a problem for variance estimation, it has the potential to make small sample sizes an issue of the past and reduce sampling error to miniscule amounts.

The remaining methodological challenge deals with the *level of detail* provided by an alternative source. Corporate data providers and vendors may be unwilling or unable to provide the level of detail BLS economic assistants collect from observation, which requires compromise and accepting aggregated data that are less than ideal for price index calculation. Their definition of a unique item may not match the BLS definition, making it difficult to price the same item over time. Limited information on product features and unstructured item descriptions require new approaches to matched model indexes and quality adjustment in the CPI. Most alternative data sources also omit sales tax information and may not provide enough information to identify the tax jurisdiction that CPI needs to apply a tax rate. In general, we have to adapt our methods on a case-by-case basis to address the specific issues of each alternative data source.

[6] Williams and Sager (2019b).

[7] Specifically the rolling year GEKS index discussed in Ivancic, Diewert, and Fox (2011)

[8] Silver and Heravi found that coefficient estimates from hedonic regressions may be affected by product cycles, which they attributed to pricing strategies including the dumping of obsolete merchandise.

[9] Sheidu (2013)

Survey Operations Challenges

Many of the challenges presented by alternative data sources for CPI deal with various aspects of survey operations, i.e. the procedures, systems, and collection strategies employed by the program. This section will highlight the primary survey operations challenges.

While velocity is often listed as one of the virtues of big data, the *time dimensions* of both corporate and secondary source datasets can be an issue – BLS timing and needs for a monthly index are not always a high priority or possible for data vendors and corporate headquarters. At times, we risk publication delays unless we are willing to truncate observations from the end of the month. In other cases, the data is only available with a lag – this is particularly the case with medical claims data as described in the Physicians and Hospitals Services case. To the extent that the CPI is making use of data from multiple sources that come in with varying lags, BLS may need to reconsider the CPI as a measure that is published and never revised, taking into consideration the impact that might have on use of the CPI for cost-of-living-adjustments and contract escalation.

BLS has control over all data processing of traditionally collected data and have many procedures and systems in place to control the overall *quality of the micro data* collected and used in CPI's outputs. With alternative data, we have to rely on others that do not always have the same data quality needs as we do. Data cleanliness can be a risk with vendor data, descriptive data isn't always collected, and data comparability over time is not guaranteed. In addition, *continuation of any vendor data source* is not guaranteed and could disappear without any warning; thus, we spend some time looking at these risks and how best to manage them. We create fallback plans, but recognize that their implementation—if needed—may not be fast or smooth enough to prevent temporary gaps in coverage in the CPI.

In order for an alternative data source to be incorporated into the aggregate CPI measure, the data must be *mapped into CPI's item categorization and geographic structure*. This is simple when a dataset's coverage directly corresponds to a CPI item category. However, in certain cases, we receive transaction data that cover a broad range of items and BLS must match items based on the company's categorizations and item descriptions; this usually involves development of concordances. The CorpX case is a prime example of this need. We have developed a machine learning system to assist in these categorizations that has greatly improved our ability to handle large datasets with hundreds of thousands of items.

Once we acquire a data source, resolve any methodological issues, and decide to incorporate it into the CPI, we must still deal with *integrating the data into our current information technology systems* that assumes data are structured according to our survey data collection process. We essentially have two ways of doing this without completely redoing all of CPI's systems. We can either replace an individual price observation in the CPI or replace a component index with an index derived from alternative data. Even in those cases, we have to decide *how best to transition* between current methods and the new alternative index – how to inform our data users, timing, addressing aggregation with other CPI components, etc. The New Vehicles case is instructive in this regard. Replacing individual price observations works well when we want to mix surveyed and alternative data in item categories. For example, we can replace the observations corresponding to one corporate respondent with alternative data while continuing to use surveyed data to represent other respondents, thus keeping outlet weights constant. However, the current system is not designed to allow us to generate new price observations, so our current strategy is to match a price or price change estimate to an existing price observation that has been selected for sampling. If we have information that cannot be matched to the existing sample (for example, a combination of seller and city that has not been selected), it cannot be used under the match and replace method. Both the Residential Telecommunications Services and New Vehicles cases are good examples of the various kinds of adaptations we are making in this regard. Ultimately, we are attempting to standardize our collection of alternative data sources to the degree possible. As we introduce alternative data into the CPI, we have to avoid a proliferation of individual respondent and secondary source systems that can only handle data from that source. Longer term we will consider more extensive changes to our IT systems to more fully utilize alternative data.

Legal, Policy, & Budgetary Challenges

Lastly, CPI needs to deal with legal, policy, and budget challenges. For secondary sources, this usually focuses on *negotiation of contracts* that are consistent with Federal laws and that meet the needs of both parties as well as making sure that costs are reasonably controllable in the longer term (there are limits on the number of option years BLS can have on a contract). Nevertheless, there is the possibility that *contract costs* can increase exponentially when it comes time for renewal and we need to plan accordingly to the extent that is possible. Sole source contracts are problematic for BLS, and without data continuity, we risk having to continually change production systems to accommodate new data and formats, which could be quite costly or lead to unpublishable indexes. In addition, for corporate data, there could be a need to enter into a formal user agreement.

The BLS's primary obstacle to *adopting web scraping* has been legal. Concerns regarding web scraping have arisen both internally and from respondents. The Confidential Information Protection and Statistical Efficiency Act is the primary US law ensuring the *confidentiality* of BLS microdata. To ensure all alternative data used in research or production is protected under CIPSEA, BLS must provide establishments, including those whose data we collect on-line, whether manually or automatically, a pledge of confidentiality promising to use the information for exclusively statistical purposes. In the case of secondary source datasets, a condition of the contract could be that the vendor be acknowledged publicly, such as J.D. Power in the New Vehicles case. In the case of web scraping, BLS cannot proceed without permission of the establishment. Moreover, Terms of Service agreements (TOS) for websites and APIs often have aspects that are problematic for Federal agencies. Website user terms and conditions often require users to agree to accept the state law in which the establishment resides rather than Federal law. Some TOS restrict storage of data, which is a requirement for CPI to ensure reproducibility. Many TOS have open-ended indemnity clauses to which Federal agencies cannot legally commit. Corporate legal departments sometimes find it simpler to refuse us data access than negotiate exemptions or alternative terms of service.

The issues related to web scraping involving private entities needed to be resolved before CPI could proceed beyond the initial research efforts. After extensive consultations with various BLS stakeholders and the DOL Solicitor, we have recently developed a policy for web scraped data in which we need to provide a pledge of confidentiality to potential website owners and obtain their consent to web scrape with the understanding that we will use best practices and, if they have a TOS, explain which terms we will not be able to follow and why. Since the data obtained via web scraping after obtaining permission will be used for exclusively statistical purposes and may be comingled with other respondent identifiable information previously collected as part of a BLS survey, we have concluded that web scraped data are subject to CIPSEA protections. Similar to the New Vehicles case, there can be situations in which web scraping involves obtaining data from a third-party vendor, such as in the Crowd-sourced Motor Fuels experience, in which we are allowed to choose to identify the vendor.

Finally, CPI has an overall goal to make sure that the transition to alternative data sources does not increase its overall budget, i.e. that this work remains at least budget neutral if not actually resulting in overall cost savings. The Food At Home case is a good example of how this emphasis on overall cost effectiveness can play out.

Experiences with Corporate Data Collection

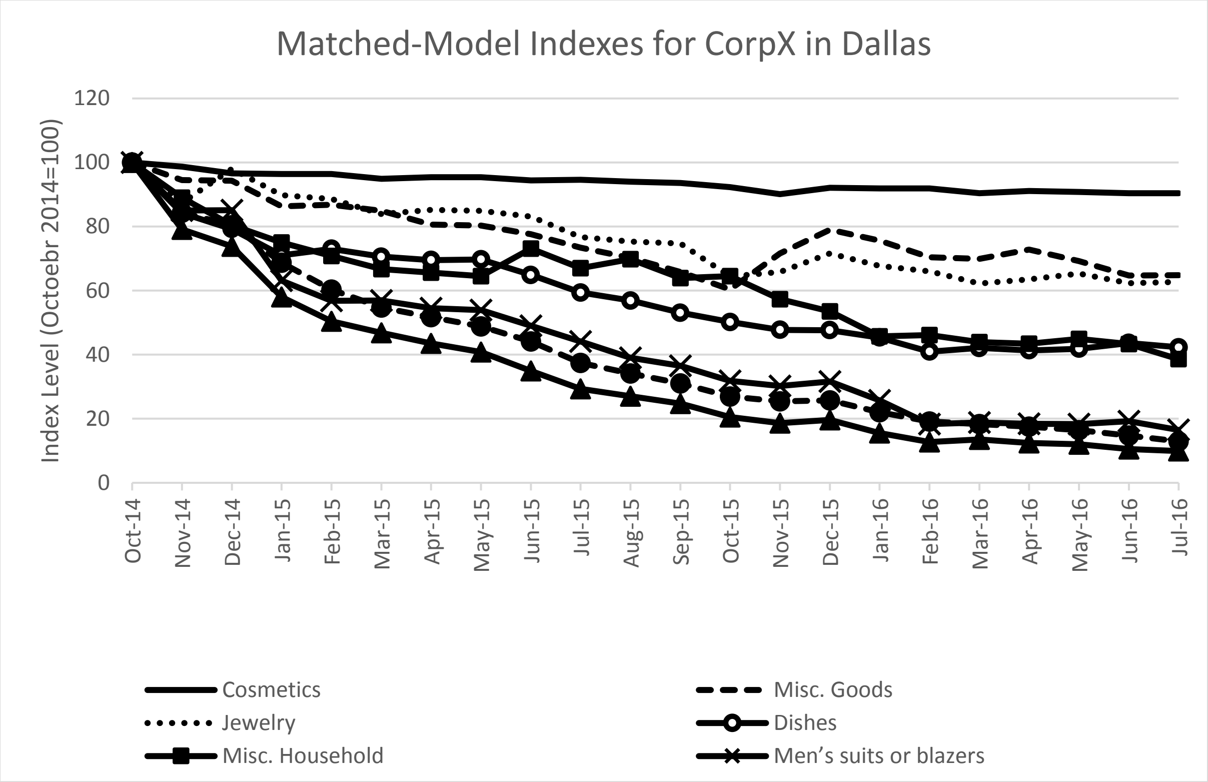
CPI's experience with corporate data sets is illustrated with two instances: one currently being used in the CPI (CorpY), and one which will be introduced in the near future (CorpX), both initially in reaction to the companies' reluctance to allow continued in-store collection. We believe these two experiences represent the variety of situations CPI can face with this type of data depending on what the company is willing to provide – not all respondents are willing to provide a complete census of their transactions for example. Our work with these two companies over a long period of time has taught the program many lessons about the ideal approach in requesting this type of data. In fact, BLS has recently been successful in expeditiously using a new proactive approach to obtain a corporate dataset of airlines fares that meets CPI's needs.

CorpX

Our experience with a department store (referred to as CorpX) illustrates the great potential of corporate datasets that contain comprehensive transactions information, but also many of the methodological and operational challenges mentioned in the previous section. The latter include dealing with limited item descriptions, product lifecycle effects in a matched-model price index, the need to map the data to the CPI item structure, the need for adjustments to CPI IT systems to accommodate the data, and the decision to pursue a short-term solution while we continue to research longer-term solutions for being able to fully realize the potential advantages that come from getting both price and real-time sales data.

In May 2016, CorpX began supplying BLS with a monthly dataset of the average price and sales revenue for each product sold for each CorpX outlet in the geographic areas covered in the CPI. (Prior to May 2016, we were obtaining data that was not approved for production use, and then CorpX restructured its database and decided to provide different data to BLS.) However, the data only includes limited descriptions of the items being sold. There is no structured data on product features, and the variable description is short and sometimes not descriptive at all. This lack of descriptive data prevents us from constructing hedonic regressions or even making informed decisions of the relative comparability of new to exiting items, limiting our ability to apply CPI's usual replacement and quality adjustment methods. The data were assessed over a period of two years for replacement of more than 1000 price quotations used in the CPI, approved for use in production, and will enter the index starting with the release of the CPI for March 2019 on April 10, 2019.

Our internal analysis has shown a tendency for matched-model indexes to drop precipitously. Several item categories show more than a 90% decline in less than two years. Most of these declines can be attributed to the pricing strategy of the retailer. Products are introduced at a high price and discounted over time. The chart below shows Tornqvist, matched-model indexes for a random selection of eight item categories in one city. Most display the largest price decline over a period of less than two years. Our findings were similar to those of Greenlees and McClelland (2010) who analyzed an earlier sample of data from the same retailer. Greenlees and McClelland also found that matched-model price indexes implied implausibly large price declines that were not remedied when treated as chain drift. They found that hedonic indexes also showed large declines unless coefficients were constrained to be a fixed value over the timespan of the estimated index.

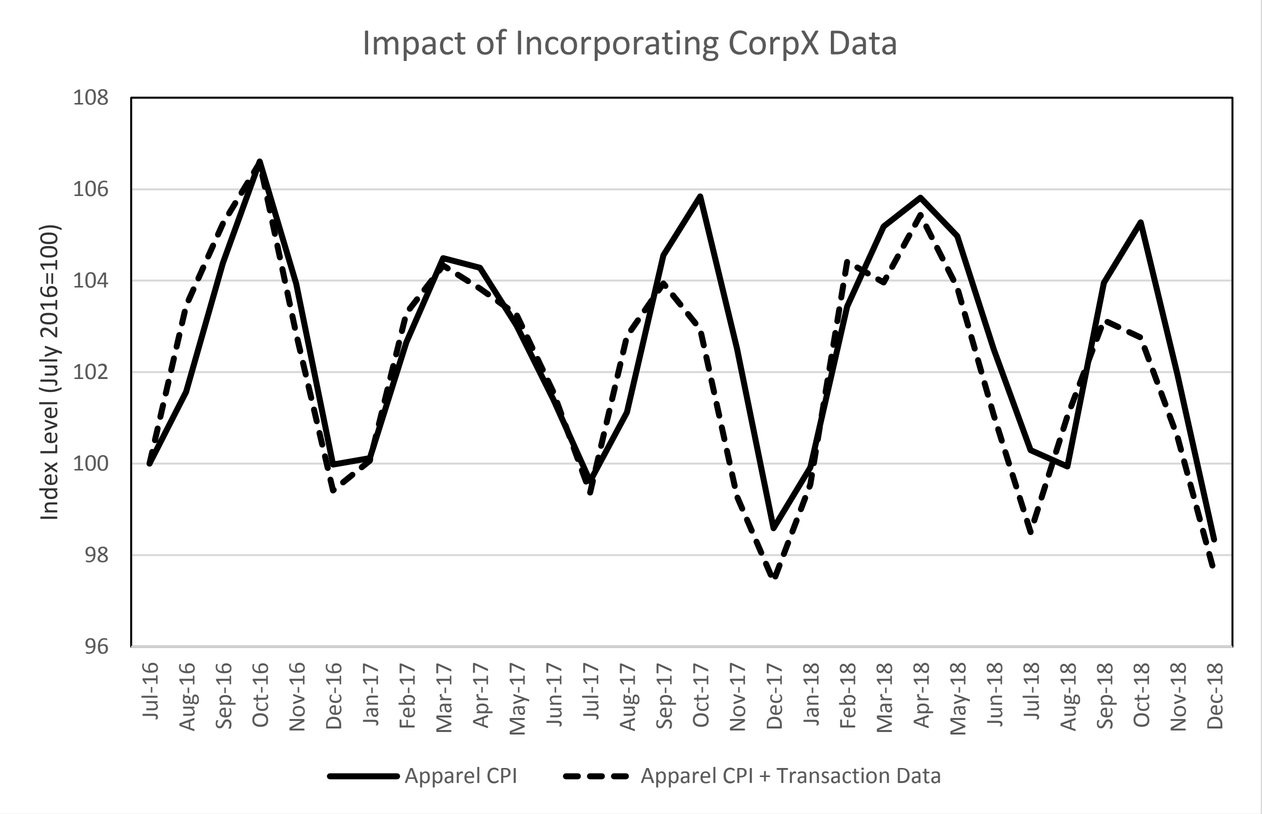


While we continue research on the best way to deal with product lifecycle effects, we have developed a short-term methodology that mimics current CPI procedures in order to begin incorporating data from this retailer into our index. Our methodology selects a sample from these transactions based on proportion of sales included in the datasets provided by CorpX and calculates match-model price relatives for these selected items over the course of a year. These matched-model indexes typically display the downward trend mentioned above. After twelve months, a new sample of products from the same item category is selected and a price relative is constructed as the average price of all new products in the item category relative to the average price of products in the category 12-months ago. This ratio between the unit prices of the new and old samples is typically positive and offsets the within-year price declines due to product lifecycles. BLS is assuming constant quality between the new sample items and the year-old sample. We are comfortable assuming this due to the merchandise and marketing strategy of the company. The sample selection process occurs twice a year corresponding to the seasonality of the items.

In order to incorporate data from CorpX into the CPI, we also had to develop a way of mapping item categorizations. The retailer provides short descriptions and categorization information for each item sold at its stores in the geographic areas covered in the CPI. Manually matching each of these items, on the order of hundreds of thousands, to a CPI item category was not feasible. Based on methods developed at the BLS for auto-coding workplace injuries,[10] we used machine learning to classify items by the CPI structure based on their descriptions. CPI staff hand-coded classifications for a segment of the items in the corporate data to create a training dataset. We then used the "bag-of-words" approach based on the frequency of word occurrences in the item descriptions. A logistic regression was then used to estimate the probability of each item being classified in each category based on the word frequency categorizations

[10] Measure, A. "Automated Coding of Worker Injury Narratives." Paper presented at JSM; Boston, MA. 2014. https://www.bls.gov/osmr/pdf/st140040.pdf

in the training data. After validating the results and reviewing low confidence predictions, BLS uses this approach with each monthly dataset to categorize new items.



The chart above compares the current published apparel price index with the experimental index that incorporates CorpX transaction data using the methodology described above. The published index does not omit CorpX entirely. Once EAs could no longer collect in stores, they collected prices for items on the store's website. The experimental index replaces these web collected prices with a price index that represents the corporately supplied data and price change from the method described above.

CorpY

Another company (referred to as CorpY) has agreed to supply BLS with prescription drug data via its corporate office. In February 2012, CorpY refused to participate in the initiation of new rotation samples due to the burden placed on in-store pharmacies. Discussions ensued between regional office staff and the company to obtain corporate data that are acceptable for CPI use and meets the confidentiality concerns of CorpY. Since March 2015, CorpY has been providing the CPI with a bimonthly dataset of average prices for a sample of their in-store prescription drug transactions.

With traditional collection methods, the CPI defines a unique item to track over time to include National Drug Code (NDC), prescription size, and insurance provider and plan or cash price. By holding these variables constant, the CPI can ensure that any price change is not due to changes in the drug's quality. The FDA-assigned NDC specifies a pharmaceutical molecule, manufacturer, and dosage. Since each NDC corresponds to a manufacturer, the CPI can also control for whether the pharmaceutical is a brand-name drug or a generic competitor. Economic Assistants (EAs) in the field collect prices for these quotes by recording list prices at prescription drug retailers. While EAs attempt to capture a realistic ratio of insurance to cash prices, the CPI is biased towards cash list prices. Respondents often refuse to provide insurance prices or simply cannot due to their database systems.

When brand-name drugs lose their patent protection and generics enter the marketplace, generic sales are slow to start due to prescriptions lasting for multiple days, weeks, or months. After approximately six months, BLS believes the generic has sufficiently penetrated the market. At this point, EAs ask pharmacists the percentage of generic versus brand-name drug sales, and based on those percentages samples brand or generic to continue pricing. If a generic is selected, the price change between brand-name and the generic is reflected in the CPI.

Ideally, CorpY would have agreed to provide a corporate dataset that provided a census of CorpY's monthly prescription transactions including a complete breakdown of brand and generic transactions. Due to the company's concerns about confidentiality and reporting burden, BLS needed to compromise and receives the bimonthly dataset mentioned above and whose features are described in more detail in the table on the next page. CorpY data defines unique items using the Generic Code Number (GCN) instead of NDC. Each GCN defines a particular drug's composition, form, and dosage strength. Unlike NDC, the GCN does not specify a manufacturer, so whether the drug is brand or generic is unknown. CorpY averages prices across brand name and generic versions. As consumers substitute between brand and generic versions of a drug, the average price will change. While there are often large differences in out-of-pocket costs between brand and generic drugs, they are generally regarded as equivalent and equally effective and thus the average price was considered acceptable by CPI.

The table on the next page compares the sampling and pricing methodology between CorpY and CPI traditional collection (called "In-Store") and demonstrates the tradeoffs and negotiations that can take place with establishments when we discuss the corporate dataset option including providing insight into how CPI evaluates the fit with its measurement objectives. In general, BLS evaluates each alternative data source to determine whether the alternative data and method used produce a result that is as good as or better than existing indexes.

Topic	CorpY	In-Store	Preference
Sampling	Probability Proportional to Size (PPS) over the past year nationally by sales excluding lowest 10% of transactions	PPS based on price of the last 20 prescriptions sold	CorpY: The last 20 prescriptions was a compromise since pharmacists were limited in their time and ability to pull data from their records. Sampling over a one-year period is likely to be more representative.
Geographic level	National	Outlet Specific	In-Store: Distribution of drug sales may differ between various regions.
Price	Average price of at least 100 transactions	Single price	CorpY: Less volatility and the switch from brands to generics is shown as a unit price change.
	Insurance prices	Mostly cash prices	CorpY: Because most consumers pay through insurance. Ideally prices would be separated by insurance plan, and CorpY averages across insurance companies and plans.
	National price	Outlet specific price	CorpY: Averaging across all stores in the US gives a more representative price at the US level, and research showed that there was little regional price variation.
	Per pill price	Per prescription price	In-Store: Per prescription price allows CPI to control for unit price differences between prescription sizes.
Patent Loss Adjustment	Unit prices by GCN average across brand and generic	Based on analyst monitoring of patents for an NDC	CorpY: Since the GCN averages across generic and branded drugs, any patent loss will be reflected in a unit price change. Monitoring patent loss is time consuming and difficult.
Timing	Bimonthly odd collection	Monthly and bimonthly odd/even collection	In-Store: CorpY only delivers data during odd months. In-store collection is done monthly or bimonthly depending on survey design.

Experiences with Secondary Data Sources

Several vendors aggregate and sell data. These datasets are typically used by marketers and are often constructed with a focus on category level sales rather than providing product level detail. Most datasets cover far more items than the CPI sample. The BLS has purchased several datasets and researched their use as replacements for production CPI components. Secondary data sources present similar issues to those found in corporate data. The data are often lacking in descriptive detail compared to information recorded by data collectors in the C&S survey. There is often a lack of transparency from secondary sources in terms of degree of willingness to fully share their methodologies with BLS. In this section, we cover five different experiences that display the variety of uses that secondary data sources can serve.

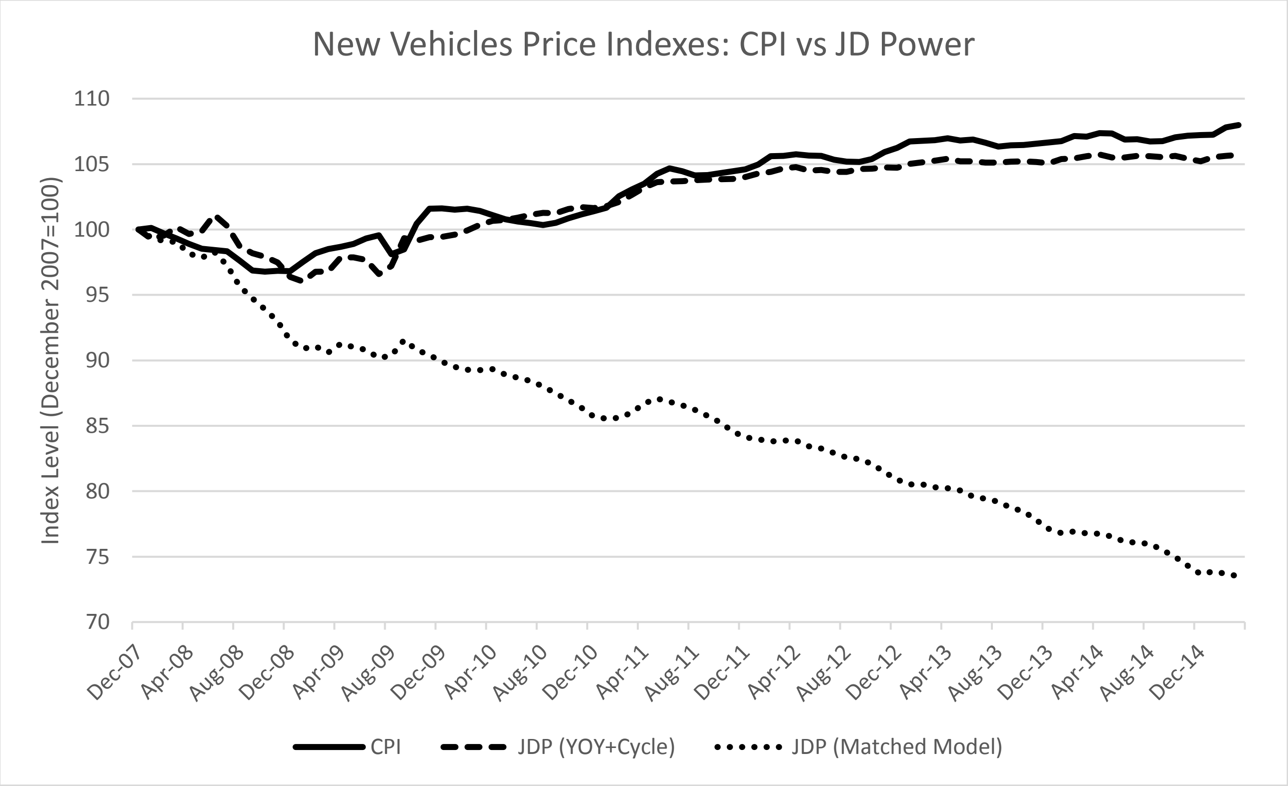
New Vehicles

In response to respondent burden, low response rates, dealer-estimated prices, and high collection costs, the BLS has pursued an alternative to its traditional data collection for new vehicles. We explored the procurement of transactions data from J.D. Power, with the vendor making its identification by BLS a condition for use of the data. In addition to addressing the issues with conventional data collection, the J.D. Power data offer higher quality information including transaction prices and real-time expenditures. The data allow a better measure of cost-of-living than the current index.

J.D. Power supplies the BLS with transaction level data that cover about one-third of the new vehicle sales in the United States. Internal analysis has shown that the market shares of vehicles in the CPI and J.D. Power's data are similar, which leads us to believe that there is little loss in representivity even though J.D. Power's dataset is not created through sampling. Each record contains information on the vehicle configuration, transaction price, and any financing set up by the dealer. The identifier available in the J.D. Power dataset that we use to define a unique item does not provide the same level of detail that we get through conventional data collection—especially the specific options sold with a given transaction.

New vehicle sales display a product lifecycle where vehicles are introduced at a high price and then discounted through the model year until they are replaced by a successor vehicle. As a result of this pattern, matched-model new vehicle price indexes show steady declines since they only reflect within-year price declines and do not account for any cross-model year price change. This index behavior may suggest chain drift, but as was the case with CorpX indexes, chain drift did not seem to be a factor since multilateral methods failed to attenuate the downward movement. Williams and Sager argue that price declines over a product's lifespan may be attributed to sellers using a price discrimination strategy that is incompatible with the assumption of stable consumer preferences in cost-of-living index theory.

Thus, as described in detail in Williams and Sager, the BLS's J.D. Power indexes use a year-over-year price relative to construct price comparisons at similar points in a vehicle's product cycle, making use of both the price and expenditure information in the dataset with a superlative index formula. Year-over-year price measurement smooths over high-frequency fluctuations in the market. In order to restore information on the short-run behavior of the new vehicle market, a monthly frequency price index is calculated. A time series filter is used to separate a cyclical component from the biased trend of the monthly frequency index. This cyclical component is combined with the year-over-year trend to create an index (YOY+Cycle) that reflects both the short- and long-term behavior of new vehicle prices.



In spring 2019, the BLS will be releasing experimental new vehicle indexes based on the methodology developed in Williams and Sager. Following a period of comment and review, the BLS may replace the new vehicles component index of the CPI with indexes based on J.D. power data. The expense of J.D. Power data is slightly less than the current cost of collecting new vehicle prices in the field, and the J.D. Power data have added benefits including a much larger sample size, transaction prices, and real-time expenditure information.

Physicians' and Hospital Services

Currently, the medical care major group has the worst response rate of all major groups in the CPI, and of that major group, "Physicians' Services" and "Hospital Services" have the highest relative importances. There are multiple reasons for this low response and all are very difficult to overcome, such as confidentiality concerns magnified by the Health Insurance Portability and Accountability Act, difficulty in determining insurance plan rates, separate physicians and billing offices, and gatekeeper issues. BLS decided to explore the feasibility of supplementing traditional data collection of cash and Medicare prices of these two items with insurance claims data. We purchased a dataset covering 2009 and 2010 medical claims data for one insurance carrier for a small sample of medical services in the Chicago metropolitan area. Average prices across all transactions for the provider/medical service combination were received each month along with the number of transactions used in creating the average price. A key research objective was to analyze the effect of using lagged insurance claims data. Claims often take months to be fully adjudicated and data processing by the vendor may take additional time. Claims data will be lagged, ranging from two to nine months, before it can be delivered to the CPI.

Indexes were calculated several ways using this dataset; the one seeming to most accurately reflect CPI methodology used a two-step weighting process. Medical services are first aggregated within outlets and weighted by their monthly quantity share to get an outlet relative. Each medical service quantity share weight is updated every month. Outlet relatives are then aggregated using outlet expenditure shares from 2008. The outlet weights were fixed for the two years of research data. Outliers were removed from the data.

Results of this preliminary research are promising but not definitive. First, BLS did not identify and request all price determining characteristics. Each medical service in Hospital Services was identified and sampled using its procedure code, i.e. Current Procedural Terminology (CPT) for outpatient and diagnosis-related group (DRG) for inpatient. Upon examination of outliers, we realized that diagnosis codes—International Classification of Diseases (ICD) codes[11]—are price determining for inpatient services in addition to the DRG. Still, price indexes created using insurance claims data tracked closely to the CPI Hospital Services index. Initial results indicate that supplementing claims data with the CPI data did not significantly change the CPI Hospital Services index values in the Chicago area, where it turns out response rates are better than average. In areas where CPI is less productive, claims data may increase accuracy.

While claims data did not significantly impact the Hospital Services index, it had a more noticeable effect on Physicians' Services. In the Chicago area, Physicians' Services price indexes combining lagged insurance claims data and CPI data for cash and Medicare prices markedly improves upon the CPI Physicians' Services index using the lagged data in the current month. Moreover, the cost of claims data is less than traditional data collection, and supplementing the CPI sample using a larger and more

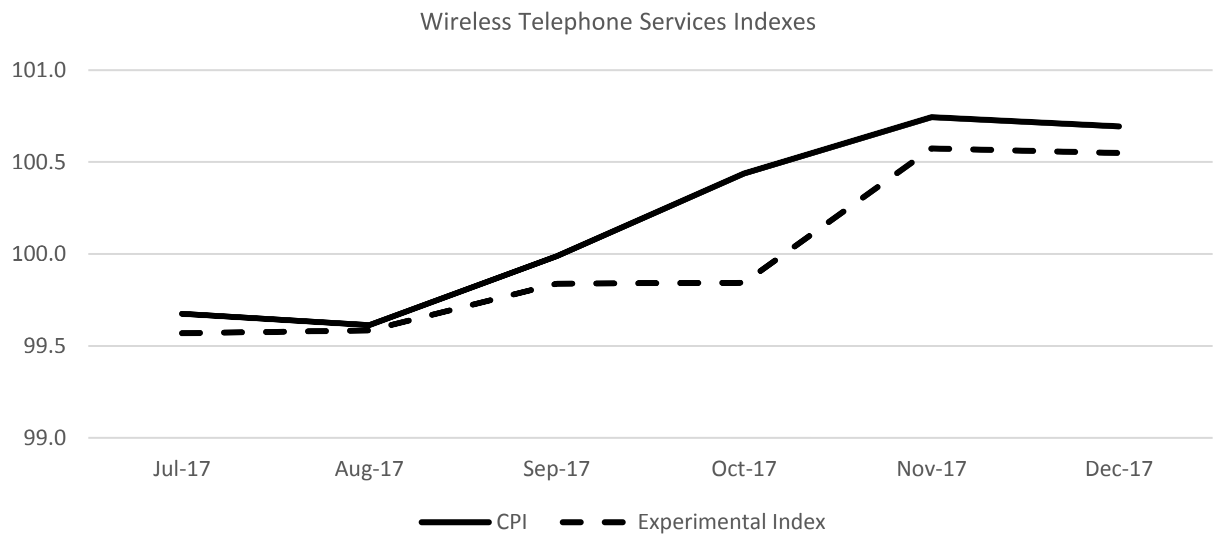
[11] The ICD is a system used by physicians and other healthcare providers to classify and code all diagnoses and symptoms.

extensive sample is highly desirable. Future plans will include expanding the research to all CPI geographic areas, a larger sample of medical services, and experimenting with time series modeling.

Wireless Telephone Services

Currently, at the request of respondents, the majority of the CPI's wireless telephone services sample is collected online using the carriers' websites. Without the assistance of a knowledgeable respondent, the CPI sample was not accurately reflecting consumer purchasing habits. The BLS prioritized the examination of alternative data for this item due to its high relative importance and online collection, and has seen promising results. Beginning in February 2018, BLS researched and leveraged a secondary source of household survey data on wireless carriers to create sampling percentages for wireless telephone services to aid field economists in selecting more representative unique items.

BLS also calculated research indexes with another secondary source that has list prices for wireless telephone service plans collected from the websites of wireless carriers. Coverage of CPI providers was over 90%. A "match and replace" methodology was used to calculate indexes, whereby the service plans in CPI collection are matched to the plan descriptions in the alternative data, the prices are replaced, and indexes recalculated using current CPI methodology and the rest of the CPI sample not covered by the data.



Over the six month period examined, the official index increased 0.69 percent while the experimental index rose 0.55 percent, a difference of 0.14 percent. This difference occurred in large part because CPI data collection is spread out over the month whereas the data in the alternative dataset were collected at one point of time in the month. Preliminary conclusions are that this data source can replicate data collected by BLS at reduced cost, with at least the same level of accuracy. We are exploring one other data source, calculating indexes over a longer period of time, and making a decision on production use in the next year. In addition, we are continuing to explore transaction price data sources.[12]

[12] On a related note, CPI started using a secondary source to assist with the process of quality adjustment for smartphones beginning with the release of January 2018 data and started directed substitution in April 2018 to bring

Residential Telecommunications Services

Similar to Wireless Telephone Services, at the request of respondents, the majority of the CPI's Residential Telecommunications Services sample is collected online using the carriers' websites. Beginning in February 2019, based on purchased household survey data, BLS created sampling percentages for landline phone service, cable and satellite television service, and internet service to aid field economists in selecting more representative unique items.

Another dataset purchase contains list prices for Residential Telecommunications services compiled by a data aggregator from several sales channels. The data to which we have access are not directly comparable to CPI prices; for example, it does not include add-on purchase like premium movie channel subscriptions or rental fees, and it includes items excluded from CPI prices such as rebates, activation, and installation. There is also no data on quantities or expenditures. To calculate experimental indexes, BLS used CPI outlet weights and distributed that weight across all items in the dataset equally. We developed matched model indexes to replicate the CPI methodology. There were significant index differences between the CPI and experimental indexes, which we determined were due to our procedures for missing data and the lack of substitution methodology. There was also difficulty in determining a unique item to price in the alternative data- what made a unique item in the dataset was not how BLS defined that item. Preliminary results show calculating the CPI for Residential Telecommunications services with alternative data is possible and with adjustments to our methods and access to a broader, richer dataset we can get results as good or of better quality than traditional field collection. Further research is planned. In addition, we are continuing to explore transaction price data sources.

Food At Home

About eight years ago, BLS purchased historical Nielsen Scantrack scanner data and used it to create indexes for comparison with the CPI Food At Home categories. The data covers five years of historical data ending in 2010 at the Universal Price Code (UPC)/geographic area, and includes some product descriptors and an average price in each observation. The Nielsen data that BLS purchased does not cover the full scope of outlet types covered in the CPI for Food at Home categories. It omits convenience stores, bakeries, butchers, smaller grocery stores, warehouse stores, and gas stations.[13] Nielsen's UPC data had to be mapped into the item categorization used in the CPI. About 80% of the UPCs could be mapped directly into a CPI category based on their Nielsen categorization, but the other 20% had to be matched manually (though we now have the experience to use machine learning to aid in mapping new items).

Initial research focused on comparing selected CPI Food At Home categories with the Nielsen Scantrak data and using the results to improve traditional data collection processes and procedures -- for example, changing the required item description forms to include different price determining characteristics. FitzGerald and Shoemaker (2013) documented some of the results of BLS's research. Later efforts turned toward exploring whether Nielsen Scantrack data could be used as replacement for certain Food at Home item categories in the CPI. The data covered around 2 million UPCs, orders of magnitude higher than the number of items tracked in the CPI. Some item categories behaved well and produced price indexes that displayed similar behavior to corresponding indexes in the CPI. Other categories displayed extreme downward declines similar to other transaction data indexes with product cycles. Researchers dealt with this downward bias by constructing common good indexes, where entering and exiting goods were

the CPI sample more in line with what consumers are purchasing. See https://www.bls.gov/cpi/notices/2017/methodology-changes.htm and https://www.bls.gov/cpi/factsheets/telephone-hardware.htm.

[13] Nielsen offered data for convenience stores, warehouse stores, and gas stations but BLS chose not to purchase that data in this initial research project.

excluded from the index. Ultimately, the BLS found that it was less expensive to collect data in stores than to pay for Nielsen Scantrack for the real-time data and geographic and outlet detail needed to support the monthly CPI. We do plan to explore whether the retailers would be willing to provide us corporate datasets, but we are not experiencing many response or collection issues in Food At Home outlets.

Housing

The CPI Housing Survey records rents from about 47,000 units selected to form a representative sample of the private rental market. A mix of property managers, renters, and their representatives are surveyed every six months. They are asked about actual, transaction rent and about what utilities and services are included in the rent along with characteristic data. BLS explored a secondary dataset of Housing rents and estimated rents to evaluate the potential for replacing or supplementing CPI Housing survey data. The secondary source dataset is not designed as a representative sample or census for a geographic area, and although it included rents and estimated rents for more than 50 million housing units, the match rate to CPI units was only about 30%. Where matches could be made, BLS matched units in the CPI sample with the same units in the dataset and calculated indexes. In the final analysis, BLS decided that the differences between CPI Housing and the secondary source dataset were too significant in terms of sample coverage and differing purposes to allow use of this secondary source in the CPI at this time. We may revisit it if the secondary source changes in the future.[14]

Experiences with Web Scraping/APIs

Currently, even when collecting information from websites, CPI data collectors manually enter it into the same data collection instrument used for in-store collection. The CPI has explored using web scraping to automate data collection from these websites instead, but encountered some legal and policy questions on which we only recently came to agreement on an approach within BLS after extensive consultations with the DOL Solicitor. Thus, to date, web scraping in CPI has only been used for research and to collect supplemental observations used in constructing hedonic models. Web scraping consists of automatically accessing a web page, parsing its contents, and recording pricing and other relevant information. Others, including MIT's Billion Prices Project, have demonstrated the benefits of using web scraping to collect massive amounts of data for the purposes of price measurement. Certain online retailers provide public access to pricing data through Application Programming Interfaces (APIs). Accessing data through APIs usually places less burden on server resources than web scraping and allows information to be collected in machine-readable format rather than parsing mark-up intended to create a human-readable webpage.

The BLS is also working on adapting its systems in order to benefit from web scraping. The BLS's current systems are highly integrated so that the variance estimation, weighting, outlet sampling, and unique item selection are all intertwined. "Plug and play", i.e. simply collecting a massive dataset of prices from the web and incorporating them into our calculations is not as straightforward as it might appear. The index calculation assumes that if three items are selected at an outlet only three items from that outlet will be used in calculations (if fewer than three are collected imputation is used). CPI will be adapting our systems in order to allow calculations when the number of prices by source varies.

[14] Although it does not involve an alternative data source, CPI management has discussed potential new modes for collecting Housing data from its respondents, including what would be the first use of the BLS Internet Data Collection Facility (IDCF) to update data for a household survey. Thus, CPI is not just looking at new data sources but at more cost effective collection modes as well.

We discuss two current research efforts related to web scraping, one using data from a crowd-sourcing website as a potential replacement for CPI's collection of motor fuels price data and one related to making our collection of airline fares from the web more cost effective. CPI is also negotiating terms of service with a person-to-person sharing app business that offered BLS use of its API.

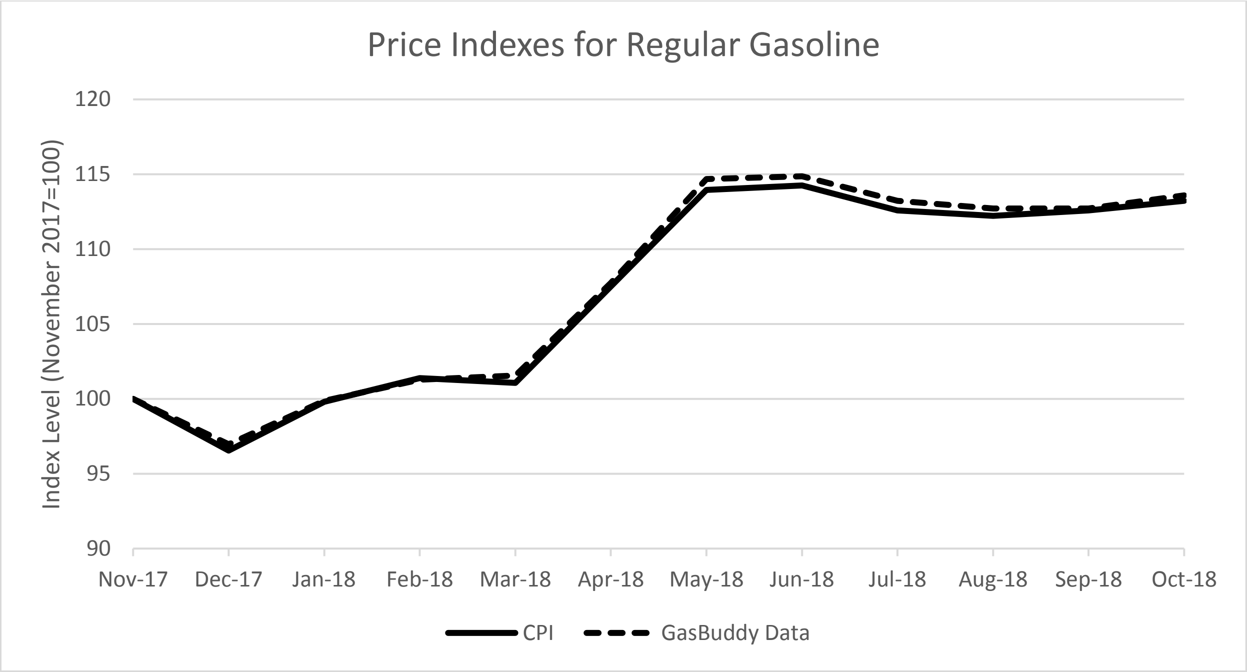
Crowd-sourced Motor Fuels

Motor fuels are one of the easier items for EAs to collect, but the large number of motor fuel outlets in the CPI (1,332 as of December 2017) leads to a large aggregate cost in terms of travel and time. Motor fuels are also an easy to define, undifferentiated product, and simple price indexes can easily be constructed with reasonable results. GasBuddy is a tech company that crowd sources fuel price collection from close to 100,000 gas stations in the US.[15] CPI obtained permission from GasBuddy to web scrape data from its website and acknowledge them as a source. We have found that indexes based on GasBuddy data closely track the CPI's gasoline indexes despite differences in weighting and details of the price information.

Much of the research work has focused on the comparisons between CPI's current data collection for motor fuels and the web scraped data. Unlike most other items in the CPI where individual item categories are sampled, all five motor fuel categories are automatically selected at any sampled motor fuel retailer in the current C&S survey. Of the five categories of motor fuels in the CPI, GasBuddy's information can replace the collected data for the three grades of gasoline and diesel, but they do not have coverage of alternative fuels. Currently, few gas stations actually offer alternative motor fuels (such as electrical charging, ethanol, E85, or biodiesel), so observations for motor fuel alternatives can be collected conventionally and comingled with the web scraped data. Currently, CPI data collectors record certain features of gasoline that may reflect how it is priced including the payment type (e.g. any cash discount or cash pricing) and whether the gasoline is ethanol free. We cannot collect this information from GasBuddy so we are ignoring possible price change due to ethanol content and the type of payment. Nevertheless, our results show that average prices and price indexes based on GasBuddy and CPI data behave very similarly, which suggests that any quality bias is not systematically large.

GasBuddy does not provide any means of weighting their price information. In incorporating GasBuddy price information into a price index, we have the choice of matching prices to the weighting information in the TPOPS survey or simply calculating an index with equal weighting for the price relatives within an area. Our most recent research has constructed indexes using the latter method and found that unweighted Jevon's price indexes based on GasBuddy data are very similar to the CPI's gasoline components despite the fact that the CPI uses TPOPS to weight gas stations. As shown in the chart at the top of the next page, the CPI showed a 13.221% increase in the price of regular gasoline over the 11 months ending in October 2018, while the GasBuddy index showed an increase of 13.595%—a difference of 0.374 percentage point. If we were to replace the current index, which is weighted, with a Jevons index, we would take advantage of the massive number of observations collected by GasBuddy (close to 175,000 price observations each month). Final evaluation of the research results and management review of recommendations for use of this source for CPI Motor Fuels outputs will occur during the remainder of this year.

[15] See https://www.gasbuddy.com/ for more information.



Airline Fares

Current pricing procedures for airline fares involve EAs in the Washington Office collecting prices from respondents' websites. Web-based pricing enables the CPI to track a defined trip month-to-month, where prices are collected by assigning each quote fixed specifications for a one-way or roundtrip fare, originating and destination cities, departure and return dates, fare class of the ticket, advance reservation, and day of the week. Each month the same advance reservation specification (designated by number of weeks) and day of the week specification will be used to collect a price. For example, a quote with a "seven week" advance reservation specification and "Tuesday" as the day of the week specification will always be priced as if the consumer booked airfare in the current month for departure seven weeks in advance on a Tuesday. This method enables the CPI to price a consistently defined trip each month in addition to accurately emulating how consumers book airfare.

In the short-term, we are researching using a match and replace methodology, meaning that we are collecting prices for each of the quotes we currently have in our sample based on the quote's specifications, and those prices are being used in our airfare sample each month. Long term, we plan to research increasing the sample size used for the calculation of price change for each respondent. We do not yet have enough collected research data to analyze the automatically collected data and associated research indexes.

BLS has begun reaching out to respondents for corporately reported data, permission to use their APIs, or permission to web scrape their sites, which is CPI's order of preference. Corporate data are preferred because we will get transaction prices and possibly weights, as well as many more price observations, but of course we will take the mode that the company is willing to provide. In the case of airfares, we have thus far received permission to web scrape from one company, essentially using the new process to gain consent described in the earlier Challenges section. We have also been receiving corporate data since October 2018 from another respondent. We plan to introduce automatic collection or corporate collection for respondents over time as each one is approved for production use.

A few words about future plans & conclusions

For over a century, the CPI has been constructed primarily using data collected by the BLS. Big data can provide information on real-time weighting, the missing fundamental piece from official price statistics for years. New alternative data sources have the potential to address many of the problems we have faced in recent years including lower response rates and higher collection costs. After several years of work on various alternative data sources, BLS has now drafted a business vision for what CPI will look like in the course of the next decade. This includes a goal to replace a significant portion of CPI direct collection with alternative data sources in the next five years. Our approach is to prioritize alternative data for item categories and outlets based on a number of factors including the relative importance of the item, the number of quotes replaced, the cost of collection, the cost of alternative data, the accuracy of the current item index, respondent relationship with BLS, ease of implementation, response rates, and the concentration of the sample for a given item. For example, fifteen establishments each account for more than 1000 price quotations apiece. We will prioritize gaining cooperation for corporate data collection from large establishments such as these as well as respondents in specific highly concentrated markets. We will also explore alternative data for item categories that may benefit in terms of accuracy and/or efficiency. As reflected in the CPI & Alternative Data Plans Table that appears in the Appendix to this paper, there are numerous items to pursue, balancing index accuracy and operating costs. The Table is organized in parallel to the CPI-U news release tables with item categories aggregated to the highest level that alternative data can be pursued.[16] The Legend at the end of the Table provides information on the contents of each column. Note that this represents only an estimate of the current plans – they are subject to change as we learn more and proceed with implementation. As of now, the Table indicates that we have either current or potential planned "experience" in item categories to some degree that represent 31.92 of the relative importance in the CPI-U.

While alternative data allow us to explore a variety of methodological improvements, our experiences to date demonstrate fundamental issues in basic indexes that require resolution. Simple techniques such as matched-model price indexes do not necessarily produce tenable results, and current CPI methods may not translate well to transaction data. We have developed ways of addressing product lifecycle with the new vehicles indexes soon to be published on an experimental basis and a short-term solution that allows us to replace manual collection of price data from the CorpX website with a corporate transaction dataset. We continue to review the academic literature for the latest transaction data price index methods while we develop new methods and/procedures for taking advantage of alternative data including several projects underway exploring econometric techniques and other ways to meet all the challenges presented by this important opportunity. We will continue to introduce alternative data incrementally in the CPI, while continuing to be mindful of core CPI measurement objectives and meeting the needs of the program's broad base of data users.

[16] For example, food at home establishments are similar enough at the lower levels that in obtaining food at home data or corporate collection, we will receive data for all item categories under food at home. In contrast, in the category Household furnishings and supplies, the outlet sample varies enough that we broke out the item categories under Household furnishings and supplies. As we seek corporate collection, there will be some overlap with other items.

References

2018. "Chapter 17. The Consumer Price Index." In *BLS Handbook of Methods*. https://www.bls.gov/opub/hom/pdf/homch17.pdf.

Cavallo, Alberto, and Ricardo Rigobon. 2016. "The Billion Prices Project: Using Online Prices for Measurement and Research." *Journal of Economic Perspectives* 30 (2): 151-178.

FitzGerald, Jenny, and Owen Shoemaker. 2013. "Evaluating the Consumer Price Index Using Nielsen's Scanner Data." JSM 2013 - Government Statistics Section, October. https://www.bls.gov/osmr/pdf/st130070.pdf.

Greenlees, J., and R. McClelland. 2010. "Superlative and Regression-Based Consumer Price Indexes for Apparel Using U.S. Scanner Data." St. Gallen, Switzerland: Conference of the International Association for Research in Income and Wealth.

Ivancic, Lorraine, W. Erwin Diewert, and Kevin J. Fox. 2011. "Scanner data, time aggregation and the construction of price indexes." *Journal of Econometrics* 161 (1): 24-35. doi:10.1016/j.jeconom.2010.09.003.

Kellar, Jeffrey H. 1988. "New Methodology Reduces Importance of Used Cars in the Revised CPI." *Monthly Labor Review* 111 (12): 34-36. http://www.jstor.org/stable/41843067.

Measure, Alexander. 2014. "Automated Coding of Worker Injury Narrative." Paper presented at JSM, Boston, MA. https://www.bls.gov/osmr/pdf/st140040.pdf.

Melser, Daniel, and Iqbal A. Syed. 2016. "Life Cycle Price Trends and Product Replacement: Implications for the Measurement of Inflation." *Review of Income and Welath*, September: 509-533. doi:10.1111/roiw.12166.

Sheidu, Onimissi. 2013. *Description of the Revised Commodities and Services Optimal Sample Design.* JSM 2013 - Government Statistics Section. October. https://www.bls.gov/osmr/pdf/st130060.pdf.

Williams, Brendan, and Erick Sager. 2018. *A New Vehicles Transaction Price Index: High-Frequency Component Extraction and a Trend Corrected Price Index.* Poster presented at: Meeting of the Group of Experts, UNECE. Geneva, May 7-9.

Williams, Brendan, and Erick Sager. 2018. *A New Vehicles Transaction Price Index: Offsetting the Effects of Price Discrimination and Product Cycle Bias with a Year-Over-Year Index.* 2018 NBER Summer Institute. Boston, MA, July. http://papers.nber.org/conf_papers/f111600.pdf.

Appendix Table: CPI and Alternative Data Plans (pending CPI Management review)

Item	RI	# quotes	concentration	issues	priority	Source of data	Experience	% sample
All items	**100.000**	**128,282**	**L**					
Food	**13.235**	**38,132**	**M**					
Food at home	7.256	32,546	M	L	M			
Food away from home	5.979	5,586	L	L				
Full service meals and snacks	2.969	1,844	L	L				
Limited service meals and snacks	2.542	2,808	M	L	M	corp	pursue	
Food at employee sites and schools	0.181	462	L	L				
Food from vending machines & mobile vendors	0.091	300	L	L				
Other food away from home	0.196	172	M	L				
Energy	**8.031**	**7,777**	**L**					
Energy commodities	4.630	4,967	M					
Fuel oil and other fuels	0.193	359	M	L				
Motor fuel	4.437	4,608	M					
Gasoline (all types)	4.344	3,778	M	L	H	scrape	20/21	100
Other motor fuels	0.094	830	M	L	H	scrape	20/21	90
Energy services	3.401	2,810	L					
Electricity	2.655	1,406	M	M	H		seek	
Utility (piped) gas service	0.747	1,404	M	M	H		seek	
All items less food and energy	**78.734**	**82,373**	**L**					
Commodities less food and energy commodities	**19.519**	**55,014**	**M**					
Household furnishings and supplies	3.336	8,479	M					
Window and floor coverings and other linens	0.258	916	H	L				
Furniture and bedding	0.883	1,881	L	L				
Appliances	0.216	610	H	L				
Other household equipment and furnishings	0.491	1,872	M	L				
Tools, hardware, outdoor equipment & supplies	0.659	1,436	H	L	M			
Housekeeping supplies	0.829	1,764	H	L	M			
Apparel	3.114	21,919	M					
Men's apparel	0.593	4,468	M	L	M	corp	some	
Boys' apparel	0.170	1,018	M	L		corp	some	
Women's apparel	1.103	8,853	M	L	M	corp	some	
Girls' apparel	0.185	2,904	M	L		corp	some	
Men's footwear	0.217	674	M	L		corp	some	
Boys' and girls' footwear	0.161	937	M	L		corp	some	
Women's footwear	0.295	1,774	M	L		corp	some	
Infants' and toddlers' apparel	0.140	580	H	L		corp	some	
Watches	0.099	303	M	L		corp	some	
Jewelry	0.152	408	M	L		corp	some	
Transportation commodities less motor fuel	6.514	0						
New vehicles	3.695	1,900	L	H	H	sec	20/21	100
Used cars and trucks	2.329	4,537	H	H	H	sec	prod	100
Motor vehicle parts and equipment	0.378	708	M	L				
Medical care commodities	1.710	5,860	H					
Medicinal drugs	1.653	5,504	H					
Prescription drugs	1.316	4,641	H	H	H	corp	some	
Nonprescription drugs	0.336	863	H	L				
Medical equipment and supplies	0.057	356	H	L				

Appendix Table: CPI and Alternative Data Plans (pending CPI Management review)

Item	RI	# quotes	concen tration	issues	pri ori ty	Source of data	Experience	% sam ple
Recreation commodities	1.792	5,835	M					
Video and audio products	0.231	1,113	H					
Televisions	0.105	350	H	L				
Other video equipment	0.027	254	H	L				
Audio equipment	0.043	265	H	L				
Recorded music and music subscriptions	0.048	244	H	L				
Pets and pet products	0.600	1,311	M	L				
Sporting goods	0.488	1,016	M					
Sports vehicles including bicycles	0.278	153	M	L				
Sports equipment	0.203	863	M	L				
Photographic equipment and supplies	0.033	272	H					
Recreational reading materials	0.113	711	M					
Newspapers and magazines	0.069	395	L	L				
Recreational books	0.044	316	H	L				
Other recreational goods	0.327	1,412	H					
Toys	0.256	958	H	L				
Sewing machines, fabric and supplies	0.023	240	H	L				
Music instruments and accessories	0.036	214	M	L				
Education and communication commodities	0.546	1,192	M					
Educational books and supplies	0.131	245	M	L				
Information technology commodities	0.415	947	H					
Personal computers & peripherals	0.315	368	H	L				
Computer software and accessories	0.024	294	H	L				
Telephone hardware, calculators, and other consumer information items	0.076	285	H	M				
Alcoholic beverages	0.963	1,243	L					
Alcoholic beverages at home	0.598	863	L	L				
Alcoholic beverages away from home	0.365	380	L	L				
Other goods	1.545	3,341	M					
Tobacco and smoking products	0.647	1,027	M					
Cigarettes	0.573	787	M	L				
Tobacco products other than cigarettes	0.059	240	M	L				
Personal care products	0.688	1,721	M					
Hair, dental, shaving, and miscellaneous personal care products	0.381	987	H	L				
Cosmetics, perfume, bath, nail preparations and implements	0.301	734	M	L				
Miscellaneous personal goods	0.210	593	H	L				
Services less energy services	**59.215**	**27,359**	**L**					
Shelter	32.893	896	M					
Rent of primary residence	7.825	N/A						
Lodging away from home	0.971	713	M					
Housing at school, excluding board	0.114	214	L	L				
Other lodging away from home including hotels and motels	0.858	499	M	L				
Owners’ equivalent rent of residences	23.723	N/A						
Tenants' and household insurance	0.374	183	H	L				
Water and sewer and trash collection services	1.079	985	L					
Water and sewerage maintenance	0.815	624	L	L				

Appendix Table: CPI and Alternative Data Plans (pending CPI Management review)

Item	RI	# quotes	concen tration	issues	pri ori ty	Source of data	Experience	% sam ple
Garbage and trash collection	0.265	361	M	L				
Household operations	0.870	605	M					
Domestic services	0.297	74	M	L				
Gardening and lawncare services	0.291	103	L	L				
Moving, storage, freight expense	0.101	373	M	L				
Repair of household items	0.105	55	L	L				
Medical care services	6.883	5,704	L					
Professional services	3.239	3,064	L					
Physicians' services	1.728	1,993	L	H	H	sec	20/21	75
Dental services	0.780	396	L	M	M			
Eyeglasses and eye care	0.316	421	L	M	M			
Services by other medical professionals	0.415	254	L	M	M			
Hospital and related services	2.591	2,640	L					
Hospital services	2.312	2,123	L	H	H	sec	20/21	85
Nursing homes and adult day services	0.191	345	L	L				
Care of invalids and elderly at home	0.087	172	L	M	M			
Health insurance	1.053	N/A				sec	prod	100
Transportation services	5.945	5,385	M					
Leased cars and trucks	0.655	265	L	H	M	sec	research	100
Car and truck rental	0.118	515	H	M	M			
Motor vehicle maintenance and repair	1.117	1,097	L	L				
Motor vehicle insurance	2.382	517	H	M	M			
Motor vehicle fees	0.539	562	L	L				
Public transportation	1.133	2,429	H					
Airline fares	0.683	1,745	H	L	M	scrape, corp	research	
Other intercity transportation	0.166	451	M	L				
Intracity transportation	0.277	233	M	L				
Recreation services	3.850	6,338	L					
Video and audio services	1.587	2,317	H					
Cable and satellite television service	1.501	1,906	H	H	H	sec	20/21	95
Video discs and other media, including rental of video	0.086	411	H	M	M			
Pet services including veterinary	0.413	265	L	L				
Photographers and photo processing	0.038	166	M	L				
Other recreation services	1.810	3,590	L					
Club membership for shopping clubs, fraternal, or other organizations, or participant sports fees	0.666	1,226	L	L				
Admissions	0.655	2,141	L	M	M			
Fees for lessons or instructions	0.217	223	L	L				
Education and communication services	6.062	5,953	M					
Tuition, other school fees, and childcare	2.900	2,566	L					
College tuition and fees	1.607	1,904	L	L				
Elementary and high school tuition and fees	0.337	220	L	L				
Child care and nursery school	0.804	268	L	L				
Technical & business school tuition & fees	0.032	174	L	L				
Postage and delivery services	0.108	461	H					

Appendix Table: CPI and Alternative Data Plans (pending CPI Management review)

Item	RI	# quotes	concentration	issues	priority	Source of data	Experience	% sample
Postage	0.094	230	H	L		sec	prod	
Delivery services	0.014	231	H	L		corp	pursue	
Telephone services	2.266	2,153	H					
Wireless telephone services	1.693	1,279	H	H	H	sec	20/21	98
Land-line telephone services	0.572	874	H	H	H	sec	20/21	95
Internet services & electronic info providers	0.780	773	H	H	H	sec	20/21	95
Other personal services	1.632	1,493	L					
Personal care services	0.623	495	L	L				
Miscellaneous personal services	1.009	998	L					
Legal services	0.304	146	L	M				
Funeral expenses	0.127	240	L	L				
Laundry and dry cleaning services	0.238	206	L	L				
Apparel services other than laundry and dry cleaning	0.029	152	L	L				
Financial services	0.240	254	M	M				

LEGEND:

RI: Relative importance as of September 2018, Consumer Price Index for All Urban Consumers: U.S. city average

# quotes: The number of quotes in our sample as of August 2018 (monthly, bimonthly odd and even)

Concentration: Percent of CPI item sample in the top ten establishments where we collect data.

- L- Less than 33% of CPI sample is in top ten establishments
- M- 33 to 66%
- H- 66 to 100%

Issues: We rated index quality issues High, Medium, and Low based on a number of factors, such as response rate, collection of list prices rather than transaction prices, collecting prices on websites due to respondent request, restricted pricing at certain times of year, difficult collection methodology, costly collection, difficult item descriptions, and the degree of subjectivity in specification descriptions. An ‘H’ means that we could substantially improve index accuracy and/or cost cost efficiency with alternative data.

Priority: Priority in seeking alternative data based on factors such as index quality issues, relative importance, size of sample, alternative data source availability. An ‘H’ means these items will be our priority to pursue, ‘M’ is next to pursue as resources are available, and a ‘blank means we currently have no plans to pursue alternative collection.

Source of data: The type of alternative data we are initially pursuing for that item category.

Scrape- web scraping or API, Corp- corporately collected data, Sec- secondary source data

Experience: The status of our alternative data progress.

- Pursue- actively pursuing one or more establishments or secondary sources
- 20/21- Items where initial research is complete and with results so far, we are expecting research to be approved for production with implementation in 2020 or 2021
- Prod: in production
- Research: actively researching alternative data
- Seek: examining alternative sources
- Some: alt data account for some percent of sample in production

% of sample: % of sample replaced either in production or based on research. Corporate blank due to disclosure protection.