This PDF is a selection from a published volume from the National Bureau of Economic Research

Volume Title: Big Data for Twenty-First-Century Economic Statistics

Volume Authors/Editors: Katharine G. Abraham, Ron S. Jarmin, Brian Moyer, and Matthew D. Shapiro, editors

Volume Publisher: University of Chicago Press

Volume ISBNs: 978-0-226-80125-4 (cloth), 978-0-226-80139-1 (electronic)

Volume URL: https://www.nber.org/books-and-chapters/big-data-twenty-first-century-economic-statistics

Conference Date: March 15-16, 2019

Publication Date: Februrary 2022

Chapter Title: Big Data in the US Consumer Price Index: Experiences and Plans

Chapter Author(s): Crystal G. Konny, Brendan K. Williams, David M. Friedman

Chapter URL: https://www.nber.org/books-and-chapters/big-data-twenty-first-century-economic-statistics/big-data-us-consumer-price-index-experiences-and-plans

Chapter pages in book: p. 69 – 98

2

Big Data in the US Consumer Price Index
Experiences and Plans

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

2.1 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 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 data for their administrative and operational needs, and some of these data can potentially be used as alternatives

Crystal G. Konny is the former Chief of the Branch of Consumer Prices at the Bureau of Labor Statistics.

Brendan K. Williams is a Senior Economist in the Branch of Consumer Prices at the Bureau of Labor Statistics.

Prior to his retirement in February 2020, David M. Friedman served as the Associate Commissioner for Prices and Living Conditions at the Bureau of Labor Statistics.

Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the view of the US Bureau of Labor Statistics. We thank Matthew Shapiro, Katherine Abraham, 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. For acknowledgments, sources of research support, and disclosure of the authors' material financial relationships, if any, please see https://www.nber.org/books-and-chapters/big-data-21st-century-economic-statistics/big-data-us-consumer-price-index-experiences-and-plans.

to BLS's surveyed data. 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 Application Programming Interfaces (APIs). Alternative data sources are not entirely new for the CPI. Starting as far back as the 1980s, 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 paper will review BLS efforts to replace elements of its traditional CPI survey with alternative data sources and discuss plans to replace and/or augment a substantial portion of CPI's data collection over the next several years.

2.2 Overview of CPI

The 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.

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.[1] Weights are derived from the reciprocal 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. BLS currently only uses a superlative index formula to produce the Chained CPI-U at the upper level of aggregation.[2] The lower-level indexes used in CPI aggregates almost all use a geometric mean index formula, which approximates a COLI under the restrictive assumption of Cobb-Douglas utility.[3]

Pricing information in the current CPI is primarily based on two surveys.

1. 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.

2. See Klick, "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.

3. For additional detail on the construction of the CPI, see "Consumer Price Index: Calculation" in the online BLS *Handbook of Methods* (https://www.bls.gov/opub/hom/cpi/calculation.htm).

BLS data collectors, known as Economic Assistants (EAs), conduct the Commodities and Services (C&S) survey by visiting 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 two surveys for price and rent data, historically in several cases CPI turned to alternative data sources, including for used cars and airline fare pricing and sales tax information.[4]

Several challenges arise in calculating the CPI using traditional data collection. First, because the CPI aims to measure 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 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. Lastly, the CPI may only be able to collect offer prices that might not reflect all the discounts applied to a transaction.

In terms of survey operations, the collection of data by BLS through pricing surveys is increasingly costly and more difficult. Metropolitan areas have generally increased in size, which causes a corresponding increase in travel costs. The growth in the number of chain stores has increased the time to obtain corporate approval to collect data. Response rates are declining as the result of 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.

2.3 Working with Alternative Data in CPI

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 could address the challenges mentioned above as

4. 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 "Measuring Price Change in the CPI: Airline Fares" for more information.

well as increase sample sizes, reflect consumer substitution patterns more quickly, reduce or eliminate respondent burden, help address nonresponse problems in the CPI's surveys, and reduce collection costs. In some instances, BLS receives real-time expenditure information as well. Data may be at a more granular level, for many more items than in the 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 has been 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 in respondent stores or on their websites. As the datasets are typically created for their own use, respondents define data elements and structure, and the BLS must adapt them to BLS systems. BLS receives varying levels of information about the datasets—in general, the information provided is what the companies are willing to give. Discussions with corporate data respondents often involve finding a level of aggregation that the corporation is comfortable providing to address their 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, who has made some effort to standardize the data elements and structure across business establishments.
3. *Web/Mobile app scraping data* are collected by BLS staff using in-house software that extracts prices and product characteristics from websites and mobile apps. 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 straightforward than maintaining web scraping code over time.

BLS needs to evaluate each alternative data source, regardless of type, to ensure it meets the measurement objectives of the CPI as well as to deal with various operational considerations. In general, CPI's process for deciding whether an alternative data source is fit for CPI use currently involves the following steps for each item or establishment:

1. Determine what item or establishment to pursue (criteria taken into consideration are reflected in the appendix, in table 2A.1)
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:
   a. match and replace individual prices in CPI with individual prices in the alternative source (see Wireless Telephone Services case)
   b. 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)
   c. replace price relatives in the CPI with estimates of price change based on new methodologies (see the CorpX and New Vehicles cases)
   d. use all establishments and items in alternative data and calculate an unweighted index (see the Crowdsourced Motor Fuels case)
6. Evaluate replacement indexes based on statistical tests and cost benefit analysis based on criteria for production use:
   a. Is the data a good fit for CPI?
   b. Is it as good as or better than current pricing methodology?
   c. Is it more cost effective or does the improvement in the index justify the additional cost?
   d. 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)

While there are numerous potential benefits to introducing new alternative data sources in the CPI as noted earlier, the CPI program has also encountered challenges that have impeded BLS from quickly incorporating alternative data into its outputs. Prior to discussing specific experiences, the paper will summarize the challenges.

2.4 Methodological Challenges

Because the CPI is designed to use its own surveyed data, BLS has encountered some challenges related to alternative data congruence with CPI methodology. The primary obstacle to dealing with transaction data in the CPI has been dealing with *product lifecycle effects*—that is, 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 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. This large increase will offset the incremental price declines over the prior product's lifecycle. While this method works in the CPI's fixed weight index, Williams and Sager (2019) found that a price comparison between exiting and new goods in a dynamically weighted index may undercorrect in situations where an exiting item is a low-inventory item on clearance, or overcorrect in other situations, and that multilateral price index methods designed to address chain drift, specifically the rolling year Gini Eltetö Köves Szulc (GEKS) index discussed in Ivancic, Diewert, and Fox (2011), did not remedy downward drift associated with product lifecycles. Greenlees and McClelland (2010) found that hedonic price indexes often exhibit the same drift as matched-model indexes. Conventional hedonic methods also do not address product lifecycle effects. Silver and Heravi (2005) 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. 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 (2016) and Ehrlich et al. (this volume).

A second obstacle relates to representativeness. 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 because of sampling, is used to allocate the CPI's sample across items and outlets to minimize variance as described in Sheidu (2013). An inaccurate estimate of variance could cause an inefficient allocation of sample.

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, and resolution may require compromise and the acceptance of aggregated data that are less than ideal for price index calculation. A corporation may define a unique item differently than BLS, making it difficult to price the same item over time. Limited information on product features and unstructured item descriptions requires 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, BLS adapts methods on a case-by-case basis to address the specific issues of each alternative data source.

2.5 Operational Challenges

While *timeliness* is often listed as one of the virtues of Big Data, it can be an issue for both corporate and secondary sources—BLS needs for a monthly index are not always a high priority or even possible for data vendors and corporate headquarters. At times, BLS risks publication delays or must accept truncating observations from the end of the month. In other cases, the data are 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 has 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, BLS has to rely on others who do not always have the same data quality needs. Data cleanliness can be a risk with vendor data, descriptive data are not 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, BLS spends some time looking at these risks and how best to mitigate them. BLS creates fallback plans but recognizes that their implementation—if needed—may not be fast enough 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 many cases, transaction data cover a broad range of items and BLS must concord these items to the CPI structure based on the company's categorizations and item descriptions. BLS developed a machine-learning system to assist in the CorpX categorizations, which has greatly improved its ability to handle large datasets with hundreds of thousands of items.

Once BLS acquires a data source, resolves any methodological issues, and decides to incorporate it into the CPI, it must still deal with *integrating the data into current CPI information technology systems*, which assume data are structured according to the traditional survey data collection process. There are essentially two ways of doing this without completely redoing all of CPI's systems—replacing an individual price observation in the CPI or replacing a component index with an index derived from alternative data. In both cases, *transition decisions* must be specified—how to inform CPI data users, timing, addressing aggregation with other CPI components, and so on. The New Vehicles case is instructive as an example of replacing a compo-

nent index with an index derived from alternative data. Replacing individual price observations works well when mixing surveyed and alternative data in item categories. For example, BLS replaces one corporate respondent's data with alternative data while using surveyed data to represent other respondents, thus keeping outlet weights constant. However, the current system is not designed to generate new price observations, so the current strategy is to match a price or price change estimate to an existing price observation that has been selected for sampling. If the alternative data include 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 made in this regard.

Ultimately, BLS must standardize collection and use of alternative data sources to the degree possible to avoid a proliferation of individual respondent and secondary source systems that can only handle data from one source. Longer term, BLS is considering more extensive changes to CPI IT systems to utilize alternative data more fully.

2.6 Legal, Policy, and Budgetary Challenges

BLS needs to deal with legal, policy, and budgetary challenges. For secondary sources, this usually focuses on *negotiation of contracts* that are consistent with federal laws and meet the needs of both parties, as well as making sure that costs are reasonably controllable in the longer term (there are limits to 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 BLS needs to plan accordingly to the extent possible. Sole source contracts are problematic for BLS, and without data continuity, the risk is having to continually change production systems to accommodate new data and formats, which could be quite costly or lead to unpublishable indexes. 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 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. The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) 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 are collected online whether manually or automatically through scraping, a pledge of confidentiality promising to use the information for exclusively statistical purposes. 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 law in the state 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 access than to negotiate exemptions or alternative terms of service.

The issues related to web scraping involving private entities need to be resolved before CPI can proceed beyond the initial research efforts. After extensive consultations with various BLS stakeholders and the DOL Solicitor, BLS recently developed a policy for web and mobile app scraped data in which BLS provides a pledge of confidentiality to potential website owners and obtains their consent to web scrape with the understanding that BLS will use best practices. TOS are negotiated to follow federal law. 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 Crowdsourced Motor Fuels experience, where CPI identifies the vendor.

Finally, CPI has an overall goal to make sure that the transition to alternative data sources *does not increase its overall budget*—that is, 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.

2.7 Experiences with Corporate Data Collection

Three companies provide corporate datasets to BLS; two are described in this section. The other company has been providing airline fare data for less than a year, so those data are still in the evaluation stage. Both companies described below initially started providing corporate data in reaction to their reluctance to allow continued in-store collection.

2.7.1 CorpX

In May 2016, a department store (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, BLS was obtaining data that were not approved for production use, and then CorpX restructured its database and decided to provide different data to BLS.) However, the data only include 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 constructing hedonic regressions or even making informed decisions on the relative comparability of new to exiting items,

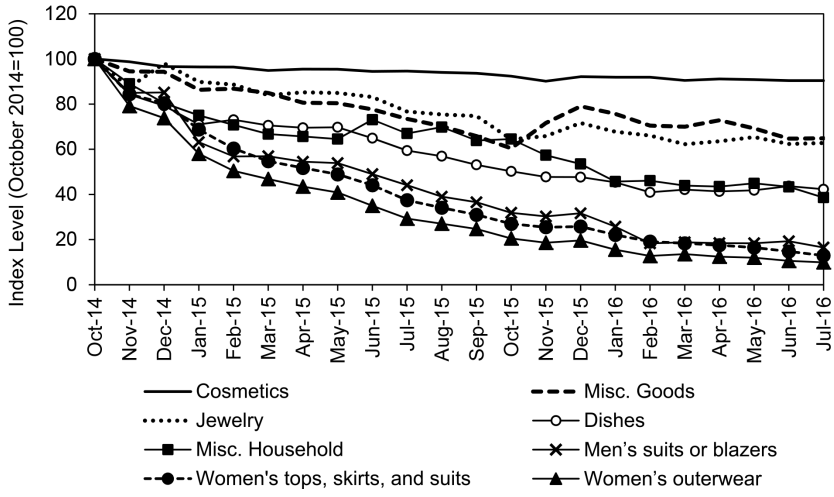

Fig. 2.1 Matched-model indexes for CorpX in Dallas
Source: CorpX data.

limiting CPI's ability to apply usual replacement and quality adjustment methods. BLS assessed the data over a period of two years for replacement of more than 1,000 price quotations used in the CPI and approved its use in production beginning with the March 2019 index.

Figure 2.1 shows Tornqvist, matched-model indexes for a selection of eight item categories in one city. Matched-model indexes drop precipitously. Several item categories show more than a 90 percent decline in less than two years. Products are introduced at a high price and discounted over time. Most indexes display the largest price decline over a period of less than two years. These findings are 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 research continues on the best way to deal with product lifecycle effects, BLS has developed a short-term methodology that mimics current CPI procedures in order to begin incorporating data from this retailer into the CPI. The methodology selects a probability proportional to size sample from sales transactions included in the dataset 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 because of product lifecycles. Since the item categories are broadly homogenous in terms of item characteristics and pricing strategy, BLS is assuming constant quality between the new sample items and the year-old sample. 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, BLS also developed 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 in Measure (2014) for auto-coding workplace injuries at BLS, CPI staff used machine learning to classify items by the CPI structure based on their descriptions, hand-coded classifications for a segment of the items in the corporate data to create a training dataset and 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 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.

Figure 2.2 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

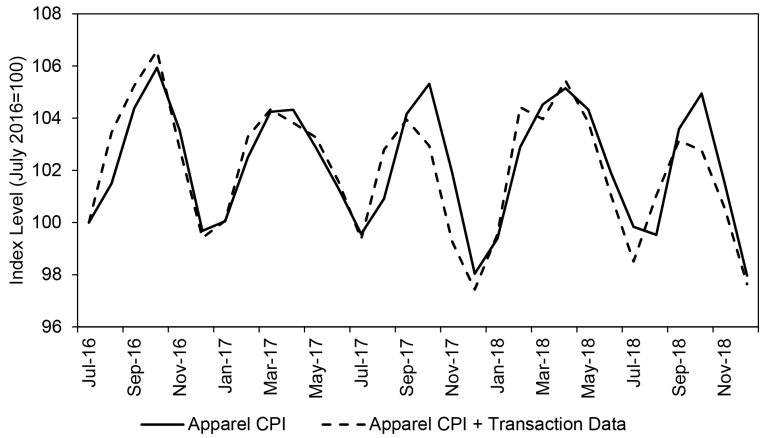

Fig. 2.2 Impact of incorporating CorpX Data
Source: CorpX data, apparel CPI.

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. The two series show similar seasonal patterns, and the inclusion of transaction data does not significantly change the index's trend.

2.7.2 CorpY

In February 2012, another company (referred to as CorpY) refused to participate in the initiation of new prescription drug rotation samples because of 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 meet 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 toward cash list prices. Respondents often refuse to provide insurance prices or simply cannot because of their database systems.

When brand-name drugs lose their patent protection and generics enter the marketplace, generic sales are slow to start as the result of 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 furnish 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 instead receives the bimonthly dataset mentioned above and whose features are described in detail in table 2.1. CorpY 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.

Table 2.1 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 discussing the corporate dataset option, including providing insight into how CPI evaluates the fit with its measurement objectives. BLS was satisfied that notwithstanding these considerations, the CorpY data are suitable for the CPI.

2.8 Experiences with Secondary Data Sources

Several vendors aggregate and sell data, both retail transaction data and offer prices. 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. Secondary sources often lack transparency in terms of degree of willingness to fully share their methodologies with BLS. In this section, we cover CPI's experience with five secondary data sources.

2.8.1 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. BLS purchases transaction-level data from J. D. Power that cover about one third of new vehicle sales in the United States. BLS analysis has shown that the market shares of vehicle makes in the CPI sample and J. D. Power's data are similar to each other and to sales data reported in industry publications, which leads to the conclusion that there is little loss of representativeness 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 item identifier available in the J. D. Power dataset does not provide the same level of detail that BLS gets 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

Table 2.1 **CorpY trade-off**

Topic	CorpY	In-store	BLS preferred approach
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 uninsured 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 price differences per pill between prescription sizes such as quantity discounts.
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	Monthly and bimonthly	In-Store: CorpY only delivers data during odd months. In-store collection is done monthly or bimonthly depending on survey design.

new vehicle price indexes show steady declines because 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 due to index nontransitivity, but as was the case with CorpX indexes, price index declines appeared to be the result of showing price decreases over a product's lifecycle. In the current BLS methodology, such declines are offset by showing a price comparison between the heavily discounted older model year and the new model year sold at or near full price. However, this method only offsets declines when using a fixed-weight index, and one of the advantages of the J. D. Power dataset is the ability to use real-time expenditure weight.

Based on the methodology developed in Williams and Sager (2019), BLS began monthly releases of a research New Vehicle index on May 15, 2019, and continues to release monthly indexes approximately three days after the release of the CPI. To construct this index, individual transaction records in the J. D. Power dataset are aggregated using a geometric mean into a unit price for a specific vehicle. Price comparisons are made between the old and new version one year apart with a year-over-year price relative to represent price change between similar points in a vehicle's product cycle. Vehicle configurations without an observed prior version are omitted. The twelfth root of these relatives are taken to represent monthly price changes, which are aggregated using the Törnqvist index with expenditure shares of each vehicle in the dataset in this month and one year ago. 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, BLS uses a time series filter to separate a cyclical component from trend in a monthly frequency index, which is susceptible to product cycle bias. The natural logarithm of this cyclical component is added to the natural log of the year-over-year trend and then exponentiated to create an index (YOY+Cycle) that reflects both the short- and long-term behavior of new vehicle prices. The YOY and YOY+Cycle indexes are compared to the CPI for New Vehicles in figure 2.3. The current BLS methodology for the CPI New Vehicles index reduces to a year-over-year price comparison since intermediate monthly price changes cancel in the fixed-weight CPI. The YOY+Cycle methodology used in the research index generalizes this measure to accommodate nonfixed weight indexes and, as a result, produces a similar measure of price change.

Following a period of comment and review, BLS may replace the new vehicles component index of the CPI with indexes based on J. D. Power data. For more detailed information, see the methodology fact sheet for the R-CPI-U-NV index on the BLS/CPI website. 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.

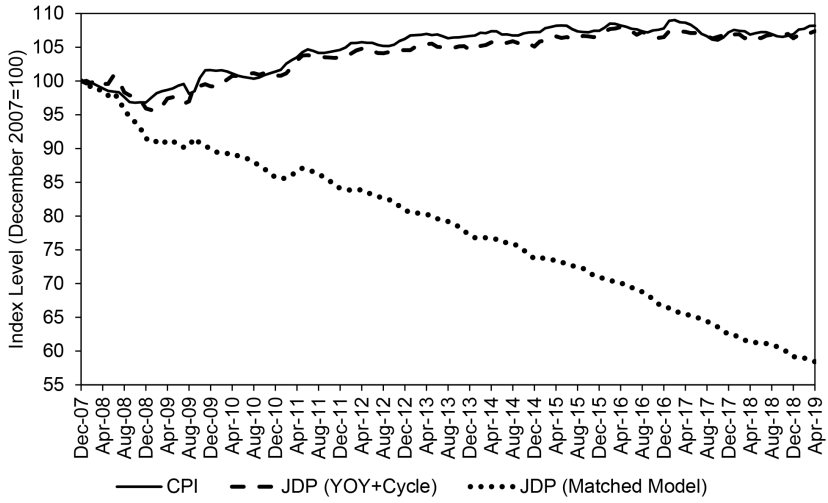

Fig. 2.3 New vehicles price indexes: CPI vs. J. D. Power
Source: JD Power, New Vehicles CPI.

2.8.2 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 importance. 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 (HIPAA), 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, and 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. BLS received average prices across all transactions for the provider/medical service combination, and 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 are lagged, ranging from two to nine months, before they can be delivered to the CPI.

BLS calculated indexes 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, the Current Procedural Terminology (CPT) for outpatient and diagnosis-related group (DRG) for inpatient. Upon examination of outliers, researchers realized that diagnosis codes—International Classification of Diseases (ICD) codes[5]—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 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, they 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 by compensating for poor response rates in surveyed data and increasing representation of insured payers. Moreover, the cost of claims data is less than traditional data collection. Future plans include expanding the research to all CPI geographic areas, using a larger sample of medical services, and experimenting with time-series modeling.

2.8.3 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 because of 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 contains list prices for wireless telephone service plans collected from the websites of wireless carriers. Coverage of CPI providers was over 90 percent.

5. The ICD is a system used by physicians and other healthcare providers to classify and code all diagnoses and symptoms. See https://www.who.int/standards/classifications/classification-of-diseases for more information.

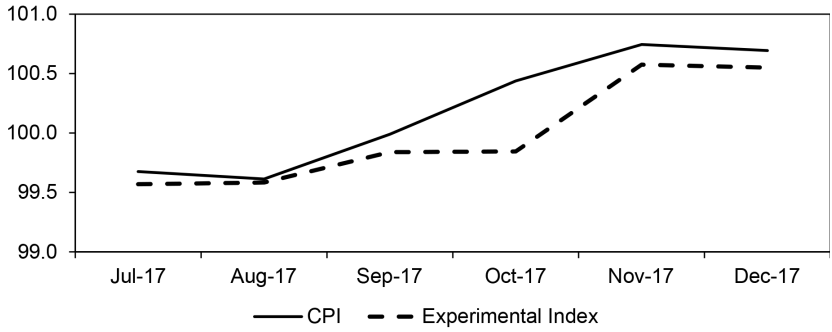

Fig. 2.4 Wireless telephone services indexes
Source: Wireless Telephone Services CPI, alternative data source.

BLS used a "match and replace" methodology 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 are recalculated using current CPI methodology and the rest of the CPI sample not covered by the data.

As reflected in figure 2.4, over the six-month period examined, the official index increased 0.69 percent while the research index rose 0.55 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. BLS preliminarily concludes that this data source can replicate data collected by BLS at reduced cost with at least the same level of accuracy. BLS is exploring one other data source, calculating indexes over a longer period of time, and will make a decision on production use in the next year, while continuing to explore transaction price data sources.[6]

2.8.4 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.

6. 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 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.

BLS purchased another dataset containing list prices for Residential Telecommunications services compiled from several sales channels by a data aggregator. The data are not directly comparable to CPI prices; for example, add-on purchases like premium movie channel subscriptions or rental fees are not included, and items excluded from CPI prices such as rebates, activation, and installation are included. There is also no data on quantities or expenditures. To calculate research indexes, BLS used CPI outlet weights and distributed that weight across all items in the dataset equally, and then developed matched model indexes to replicate the CPI methodology. There were significant index differences between the CPI and research indexes, which researchers determined were due to procedures for missing data and the lack of substitution methodology in the research index series. There was also difficulty in determining a unique item to price in the alternative data. Nevertheless, preliminary results demonstrate that it is possible to calculate the CPI for Residential Telecommunications services with alternative data. With access to a broader, richer dataset, BLS can get results with as good or better quality than traditional field collection. Thus, further research is planned in addition to exploring transaction price data sources.

2.8.5 Food at Home

BLS purchased historical Nielsen Scantrack scanner data and used it to create indexes for comparison with the CPI Food at Home categories. The purchased dataset 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 for each observation. The Nielsen data that BLS purchased do not cover the full scope of outlet types covered in the CPI for Food at Home categories, omitting convenience stores, bakeries, butchers, smaller grocery stores, warehouse stores, and gas stations.[7] BLS mapped Nielsen's UPC data into the item categorization used in the CPI. About 80 percent of the UPCs could be mapped directly into a CPI category based on their Nielsen categorization, but the other 20 percent had to be matched manually (though BLS now has experience using 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, improving the price-determining characteristics on data collection forms to better measure quality change. Later efforts, including work documented in FitzGerald and Shoemaker (2013), 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

7. Nielsen offers data for convenience stores, warehouse stores, and gas stations but BLS chose not to purchase those data in this initial research project.

higher than the number of items tracked in the CPI. Some item categories produced price indexes similar to corresponding CPIs. Other categories with product cycles displayed extreme downward declines similar to other transaction data indexes. Researchers dealt with this downward bias by constructing common goods indexes, where entering and exiting goods were excluded from the index. Ultimately, BLS found that it was less expensive to collect data in stores than to pay for Nielsen Scantrack real-time data and the geographic and outlet detail needed to support the monthly CPI. BLS plans to explore whether retailers would be willing to provide us corporate datasets, but unlike the examples discussed above, BLS has not yet experienced many response or collection issues in Food at Home outlets.

2.8.6 Housing

The CPI Housing Survey records rents from about 47,000 units selected to form a representative sample of the private rental market. Every six months a mix of property managers, renters, and their representatives are asked about actual transaction rent and 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 percent. Rents in the secondary source appeared much more volatile than those in the CPI, in part because the CPI includes ongoing and renewed leases while the secondary source estimates the current market rate for new rentals. 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 use this secondary source in the CPI at this time. BLS is exploring alternative data sources.[8]

2.9 Experiences with Web Scraping/APIs

Currently, even when collecting information from websites, CPI data collectors manually enter data into the same data collection instrument used for in-store collection. The CPI is exploring using web scraping to automate data collection from these websites instead, given recent agreement on an acceptable approach within BLS after consultation with the DOL solici-

8. 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.

tor. 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 (Cavallo and Rigobon 2016), 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 APIs, which 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 variance estimation, weighting, outlet sampling, and unique item selection are all intertwined. "Plug and play," simply collecting a massive dataset of prices from the web and incorporating them into CPI calculations, is not as straightforward as it might appear. The index calculation system assumes that a fixed number of observations are selected from each respondent. For example, if three unique products are selected at an outlet, only prices from these three observations will be used in calculations from that outlet. (When an observation cannot be collected, imputation is used.) If this respondent gave us a corporate dataset of thousands or millions of observations, our systems would not be able to accommodate additional observations beyond the three that had been selected for sampling. CPI will be adapting its systems in order to allow calculations when the number of prices by source varies.

We discuss two current research efforts related to web scraping—one using data from a crowdsourcing website as a potential replacement for CPI's collection of motor fuels price data and one related to making BLS 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.

2.9.1 Crowdsourced 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 high aggregate cost in terms of travel and time. Motor fuels are also an easy to define, undifferentiated product. GasBuddy is a tech company that crowdsources fuel price collection from close to 100,000 gas stations in the US.[9] CPI obtained permission from GasBuddy to web scrape data from its website and acknowledge them as a source.

Unlike most other items in the CPI where individual item categories are sampled, all five motor fuel categories are automatically selected at any

9. See https://www.gasbuddy.com/ for more information.

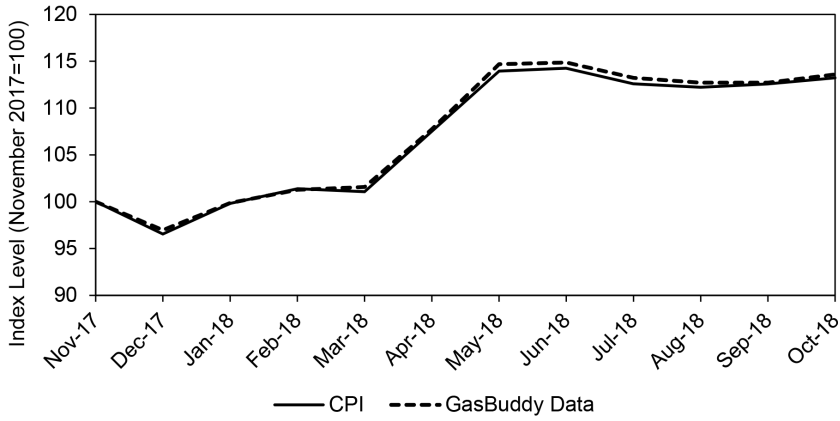

Fig. 2.5 Price indexes for regular gasoline
Source: Gasoline, unleaded regular CPI, GasBuddy.

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. CPI data collectors record certain features of gasoline that may affect pricing, including the payment type (e.g., any cash discount or cash pricing) and whether the gasoline is ethanol free, both characteristics not available from GasBuddy.

GasBuddy does not provide any means of weighting their price information. In incorporating GasBuddy price information into a price index, CPI had 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. BLS 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. Figure 2.5 shows results from one of the gasoline indexes. The CPI showed a 13.221 percent increase in the price of unleaded regular gasoline over the 11 months ending in October 2018, while the GasBuddy index showed an increase of 13.595 percent—a difference of 0.374 percentage points. Preliminary 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. BLS is calculating average prices and indexes over a longer period of time and will evaluate results and make a decision on production use in the next year.

2.9.2 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, one-way or roundtrip fare, originating and destination cities, departure and return dates, and fare class of the ticket. 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.

BLS is negotiating with respondents for corporately reported data, permission to use their APIs, or permission to web scrape their sites, which is CPI's order of preference. BLS prefers corporate data for transaction prices and possibly weights, as well as many more price observations, but will accept the mode the company is willing to provide. BLS received permission to web scrape from one company and has also been receiving corporate data since October 2018 from another respondent.

In the short term, research is focusing on a match and replace methodology, meaning collection of prices for each of the quotes currently in the CPI sample based on the quote's specifications. Long term, research will focus on increasing the sample size used for the calculation of price change for each respondent. BLS has not yet collected enough data to analyze the automatically collected data and associated research indexes. The plan is to introduce automatic collection or corporate collection for respondents over time as each one is approved for production use.

2.10 A Few Words about Future Plans and Conclusions

For over a century, the CPI has been constructed primarily using data collected by BLS staff. 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 faced in recent years, including lower response rates and higher collection costs. After several years of work on various alternative data sources, BLS now has a goal to replace a significant portion of CPI direct collection with alternative data sources by 2024. BLS will 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, 15

establishments each account for more than 1,000 price quotations apiece. BLS will prioritize gaining cooperation for corporate data collection from large establishments such as these, as well as respondents in specific highly concentrated markets, and will also explore alternative data for item categories that may benefit in terms of accuracy and/or efficiency. In addition to agency-specific efforts, the BLS is working with our partner statistical agencies to collaborate on acquiring new data sources and exploring new uses for existing administrative and survey datasets. BLS is also sponsoring a new National Academy of Sciences Committee on National Statistics (CNSTAT) panel, comprised of academic and other experts, to investigate three key methodological issues, including how BLS should modify current CPI methodology to deal with the challenges presented by blending traditionally collected data with alternative data.[10]

As reflected in table 2A.1 at the end of 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 at which alternative data can be pursued. The legend at the end of the table provides information on the contents of each column. This is the current plan and will change as progress is made and experience gained. As of now, table 2A.1 indicates that the CPI has either current or potential planned "experience" in item categories to some degree, representing about 30 percent of the relative importance in the CPI-U.

While alternative data usage may result in a variety of methodological improvements, research to date demonstrates fundamental issues 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. BLS has developed ways of addressing product lifecycle with the research new vehicles indexes, and a short-term solution that allows us to replace manual collection of price data from the CorpX website with a corporate transaction dataset. BLS continues to review the academic literature for the latest transaction data price index methods, while developing new methods and procedures for taking advantage of alternative data and the challenges presented by this important opportunity. BLS will continue to introduce alternative data incrementally in the CPI, while being mindful of core CPI measurement objectives and meeting the needs of the program's broad base of data users.

10. See https://www.nationalacademies.org/our-work/improving-cost-of-living-indexes-and-consumer-inflation-statistics-in-the-digital-age for more information, including links to the materials from the two public meetings on October 7 and October 30, 2020, regarding uses of alternative data for consumer price indexes at BLS and elsewhere.

Appendix

Table 2A.1 Preliminary CPI alternative data plans

Item	RI	# Quotes	Concentration	Issues	Priority	Source of data	Experience	% sample
All items	100.000	128,282	L					
	0	2						
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					
Fuel oil and other fuels	0.193	359	M	L				
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
Electricity	2.655	1,406	M	M	H		seek	
Utility (piped) gas service	0.747	1,404	M	M	H		seek	
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			

(*continued*)

Table 2A.1 (cont.)

Item	RI	# Quotes	Concentration	Issues	Priority	Source of data	Experience	% sample
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	7,145						
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					
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				
Recreation commodities:	1.792	5,835	M					
Video and audio products	0.231	1,113	H					
Pets and pet products	0.600	1,311	M	L				
Sporting goods	0.488	1,016	M					
Photographic equipment and supplies	0.033	272	H					
Newspapers and magazines	0.069	395	L	L				
Recreational books	0.044	316	H	L				
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				
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					
Other goods:	1.545	3,341	M					
Tobacco and smoking products	0.647	1,027	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				
Shelter:	32.893	896	M					
Rent of primary residence	7.825	N/A						
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						
Water and sewerage maintenance	0.815	624	L	L				
Garbage and trash collection	0.265	361	M	L				
Household operations	0.870	605	M					
Medical care services:	6.883	5,704	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 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

(*continued*)

Table 2A.1 **(cont.)**

Item	RI	# Quotes	Concentration	Issues	Priority	Source of data	Experience	% sample
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				
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					
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				
Club membership for shopping clubs, fraternal, or other organizations. . . 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					
Postage	0.094	230	H	L		sec	prod	
Delivery services	0.014	231	H	L		corp	pursue	
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					

Notes:

RI: Relative importance as of September 2018, Consumer Price Index for All Urban Consumers: US city average. Subcategories may not add up to the RI at the category level because of unsampled items not displayed on this table.

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

Concentration: Percent of CPI item sample in the top ten establishments where data is collected.

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

Issues: 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 BLS could substantially improve index accuracy and/or 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 BLS priority to pursue, 'M' is next to pursue as resources are available, and a blank means BLS currently has no plans to pursue alternative collection.

Source of data: The type of alternative data initially pursued for that item category. Scrape: web scraping or API; Corp: corporately collected data; Sec: secondary source data

Experience: The status of BLS 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, BLS is 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.

References

Bureau of Labor Statistics. 2020. "Consumer Price Index." In *Handbook of Methods*. https://www.bls.gov/opub/hom/cpi.

Cavallo, Alberto, and Ricardo Rigobon. 2016. "The Billion Prices Project: Using Online Prices for Measurement and Research." *Journal of Economic Perspectives* 30 (2): 151–78.

FitzGerald, Jenny, and Owen Shoemaker. 2013. "Evaluating the Consumer Price Index Using Nielsen's Scanner Data." Paper presented at the Joint Statistical Meetings 2013—Government Statistics Section, Montreal, Canada, October 2013. https://www.bls.gov/osmr/research-papers/2013/pdf/st130070.pdf.

Greenlees, J., and R. McClelland. 2010. "Superlative and Regression-Based Consumer Price Indexes for Apparel Using U.S. Scanner Data." Paper presented at the Conference of the International Association for Research in Income and Wealth, St. Gallen, Switzerland, August 27, 2010.

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. https://doi.org/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.

Klick, Joshua. 2018. "Improving Initial Estimates of the Chained Consumer Price Index." *Monthly Labor Review* (February). https://doi.org/10.21916/mlr.2018.6.

Measure, Alexander. 2014. "Automated Coding of Worker Injury Narrative." Paper presented at the Joint Statistical Meetings 2014—Government Statistics Section, Boston, MA, August 2014. https://www.bls.gov/osmr/research-papers/2014/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 Wealth* 62 (3): 509–33. https://doi.org/10.1111/roiw.12166.

Sheidu, Onimissi. 2013. "Description of the Revised Commodities and Services Optimal Sample Design." Paper presented at the Joint Statistical Meetings 2013—Government Statistics Section, Montreal, Canada, October 2013. https://www.bls.gov/osmr/research-papers/2013/pdf/st130060.pdf.

Silver, Mick, and Saeed Heravi. 2005. "A Failure in the Measurement of Inflation: Results from a Hedonic and Matched Experiment Using Scanner Data." *Journal of Business and Economic Statistics* 23 (3): 269–81. www.jstor.org/stable/27638820.

Williams, Brendan, and Erick Sager. 2019. "A New Vehicles Transaction Price Index: Offsetting the Effects of Price Discrimination and Product Cycle Bias with a Year-over-Year Index." Bureau of Labor Statistics Working Papers, No. 514. https://www.bls.gov/osmr/research-papers/2019/pdf/ec190040.pdf.