

Overview of BLS CPI alternative data initiative

Anya Stockburger

Chief, Branch of Revision Methodology

Consumer Price Index Division

prepared for
National Academies of Sciences, Engineering, and Medicine
Committee on National Statistics
Virtual Meeting with BLS
October 7, 2020



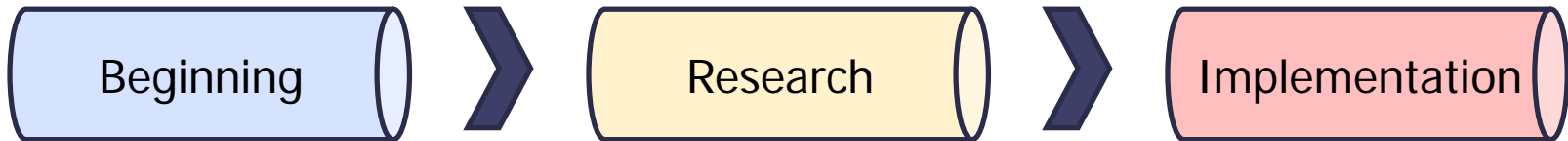
Alternative Data: current status

- **Goals:** improve accuracy of elementary indexes; improve efficiency of data collection
- **Strategic Objective:** convert a significant proportion of market basket from traditional collection to non-traditional sources and collection modes by 2024
- **Current Status:** 3.4% of market basket

| Category | Data Source | Implementation Notes |
|--------------------------|---------------------------|-----------------------------------|
| Apparel, household goods | Corporate data | Implemented March 2019 |
| Prescription drugs | Corporate data | Implemented May 2016 |
| Postage | Publically available data | Oldest use of “alternative” data |
| Used cars | Purchased data | Another long-time alt data source |

- **Other alternatives under investigation:** Respondent self-reporting, API, web-scraping, additional corporate sources or data acquisition via third party vendors

Alternative Data Pipeline: Current Projects (18%)



5 Projects = 2%

Collecting data

- Hotels
- Apparel
- General merch
- Food
- Housing

6 Projects = 9%

- Medical
- Wireless phone
- Residential telecomm services
- Airfare
- Vehicle leasing
- College tuition

3 Projects = 7%

- Motor fuels
- Airline
- New vehicles


Beginning Phase

- Initiative goals:
 - ▶ Identify continuous flow of new projects
 - ▶ Expand field expertise in new cooperation requests and collection methods
- Challenges overcome/in-progress:
 - ▶ Legal: better understanding of legal requirement for web-scraping/API access
 - ▶ Automated data collection/transfer: Developed options (BLS internet data collection facility, DMZ server), beginning work on production-grade web-scraping
 - ▶ Cooperation tool-kit: Developed materials for field staff alt data requests



Data Request Negotiations

Data Granularity

| Data Granularity | | Sales data (price and quantity sold, in preference order) | Item Coverage (in preference order) | Outlet Coverage (in preference order) | Time Coverage (in preference order) |
|---|---|---|---|---|---|
|  | A | Unique Item (UI) by price point by outlet | <ol style="list-style-type: none"> All items sold. Sample of Items > CPI Sample Items in the CPI Sample. Sample of Items < CPI Sample | <ol style="list-style-type: none"> All U.S. outlets in the chain. All outlets in the CPI PSU sample and non-self-representing PSUs not selected for the CPI sample. All outlets in CPI PSUs. All outlets in the CPI Sample. | <ol style="list-style-type: none"> Pricing period averages Monthly averages One day in each of 3 pricing periods in the month. |
| | B | UI by specific outlet | | | |
| | C | UI by City/PSU | | | |
| | D | Item category by specific outlet | | | |
| | E | Item category by City | | | |
| | F | Unique Item by region or national data | | | |
| | G | Item category by region or national data | | | |

Unit Level Aggregation

■ Defining a unique item

- ▶ Hotel = property, check-in date, length of stay, day of the week, advance reservation, occupancy
- ▶ American Hotel and Lodging Association: 54,000 properties, 5 million guestrooms, 1.1 billion guest nights annually

■ Defining an item category

- ▶ Corporate categorizations not for statistical purposes
- ▶ Categories might not be broadly homogeneous (quality issues)
- ▶ Categories might not map neatly back to CPI categories

Research

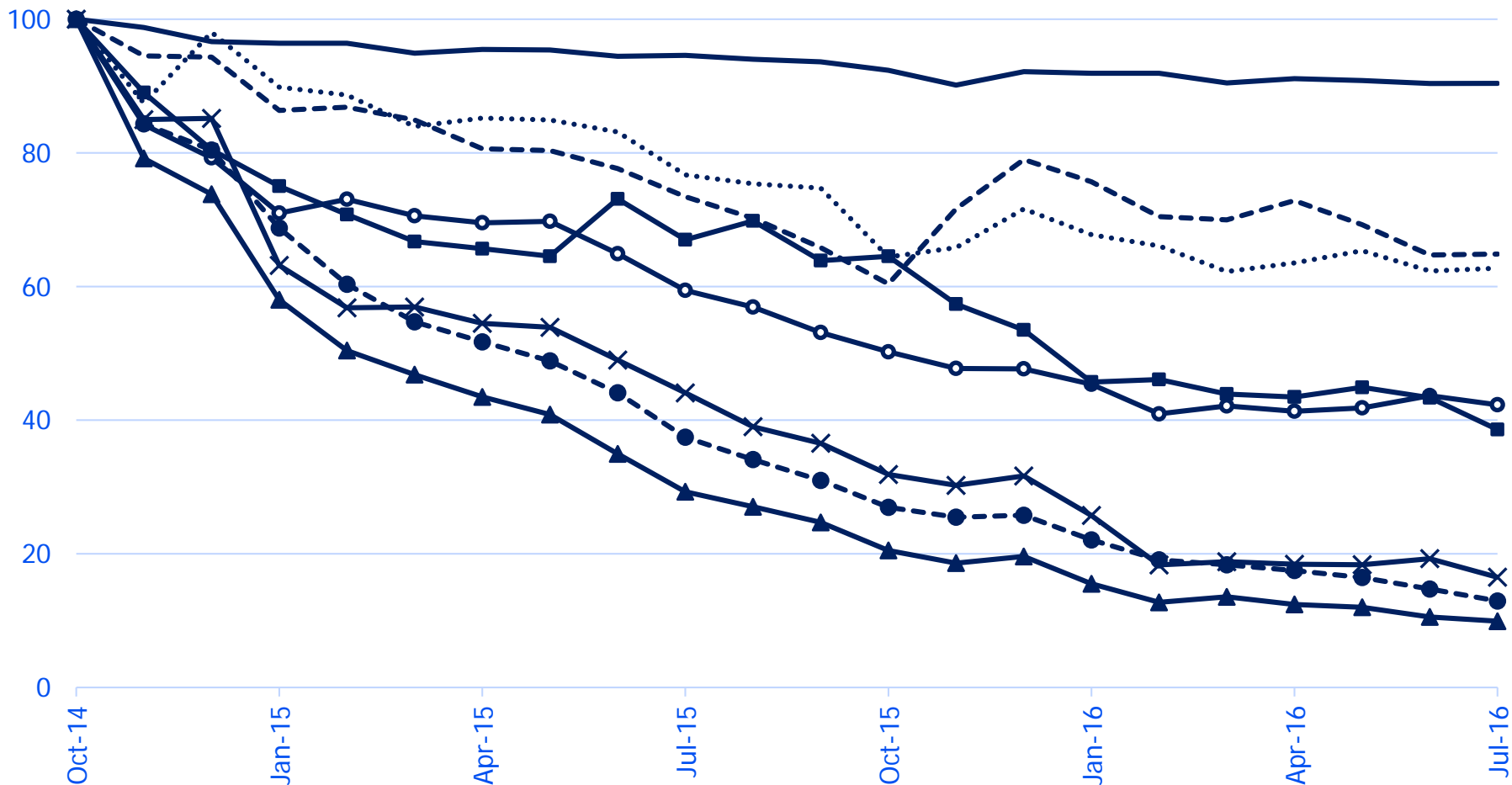
■ Initiative goals

- ▶ Standardize methods and streamline research to have a continuous flow of recommendations and improvements
- ▶ Expand researcher expertise in new price index methods

■ Challenges overcome/in-progress

- ▶ Data quality metrics: project level (when to approve?), index quality (response rates and variances enough?)
- ▶ Methodology: coverage/representativeness, aggregation issues (substitution bias vs chain link bias), product lifecycle issue, treatment of quality change, blending issue

Product Lifecycle Effect (CorpX)



U.S. BUREAU OF LABOR STATISTICS • bls.gov



Blending problem: Current

Aggregate
cells

Dresses
US

Lapsyeres
CE weights

Elementary
item/area
cells

Dresses
Wash, DC

Geomeans
Sampling Weights

Survey
Price
Quotes

#1-2

#3-4

#5-6

CorpX

Data collection

Tablet

FTP

Blending problem: Future

Aggregate
cells

Dresses
US

Laspeyres
CE weights

Elementary
item/area
cells

Dresses
Wash, DC

Geomeans
Weights?

Item/area
Substrata

Survey

Corp20

Corp30

Survey
Price
Quotes

#1-2

#3-4

#5-6

Price
Observations

Corp10

CorpX

Corp20

Corp30

Data collection

Web-
scraping

Tablet

FTP

Implementation

■ Initiative goals

- ▶ Build infrastructure to provide insertion points for non-traditional data collection (survey price quote and index relative)
- ▶ Streamline development of price relative calculation for item/area substrata
- ▶ Identify long-term system architecture

■ Challenges overcome/in-progress:

- ▶ System complexity: support family of indexes and products (average prices, “special” relatives and aggregates)
- ▶ Publication timeline: 5-7 days to receive and process; indexes are final upon release (no revision); must have back-ups identified



Contact Information

Anya Stockburger

stockburger.anya@bls.gov

