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

Crystal Konny, Brendan Williams, and David Friedman

Federal Economic Statistical Advisory Committee Meeting - June 14, 2019
Potential Benefits

- Transaction prices
- Larger sample sizes
- Reduced collection costs
- Reduced or eliminated respondent burden
- Data descriptiveness
- Real-time expenditures and weights
Challenges

- Methodological
  - Product life cycle, representativeness, data descriptiveness

- Operational
  - Data lag, continuity, quality verification
  - Geographic structure
  - System design

- Legal, Policy, and Budgetary
  - Contracting for data, webscraping agreements, confidentiality concerns
Alternative Data

- Data not collected through traditional field collection procedures by BLS staff
  *(traditional = in-store/on-phone/manually on website)*

- Three main categories:
  - Corporate
  - Secondary Source
  - Web/Mobile app scraping data

- Decade of explorations & pilot projects – transition into production
General steps for Alt Data Projects

- Determine what to pursue
- Evaluate options
- Evaluate selected source (definition, coverage, other quality dimensions)
- Evaluate data quality over predefined time
- Methods to test
- Evaluate results
- Transition to production?
Criteria for use in production (to date)

- As good or better than current pricing methodology
- Does improvement in index justify any additional costs – cost effective?
- In general, is it a good fit for CPI?
- Use of short-term solution while continuing to research longer-term improvements
Corporate Data
Impact of Incorporating CorpX

Apparel CPI

Apparel CPI + Transaction Data
CorpY

- February 2012 refused to initiate new prescription drug sample
- March 2015 agreement to supply data corporately
- May 2015 first use in index
<table>
<thead>
<tr>
<th><strong>Item Selection</strong></th>
<th>CorpY</th>
<th>In-store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability Proportional to Size (PPS) over the past year nationally by sales excluding lowest 10% of transactions</td>
<td>PPS based on price of the last 20 prescriptions sold</td>
<td></td>
</tr>
<tr>
<td><strong>Geography</strong></td>
<td>National</td>
<td>Outlet Specific</td>
</tr>
<tr>
<td><strong>Price</strong></td>
<td>Average price of at least 100 transactions</td>
<td>Single price</td>
</tr>
<tr>
<td>Insurance prices</td>
<td>Mostly cash prices</td>
<td></td>
</tr>
<tr>
<td>National price</td>
<td>Outlet specific price</td>
<td></td>
</tr>
<tr>
<td>Per pill price</td>
<td>Per prescription price</td>
<td></td>
</tr>
<tr>
<td>Patent Loss</td>
<td>Unit prices averaged across brand and generic</td>
<td>Based on analyst monitoring of patents for an NDC</td>
</tr>
<tr>
<td>Data Frequency</td>
<td>Bimonthly odd collection</td>
<td>Monthly and bimonthly odd/even collection</td>
</tr>
</tbody>
</table>
Secondary Source Data
Hospitals and Physicians’ Services

- Relative Importance 4.04%; response rate for Medical Care is 48.1%
- 4,116 price quotes
- Cash price overrepresented
- High respondent burden
- High collection costs
- Difficult collection methodology
- Researching use of medical claims datasets
New Vehicle Observations

Number of Observations/Month

0 100,000 200,000 300,000 400,000 500,000

CPI  JDPower
Experimental Index for New Vehicles

Exp New Vehicles (Untaxed)  Official New Vehicles (Taxed)
Web/Mobile app scraping data
Plans
Establishing Priorities

- Relative importance of the item
- Number of quotes replaced
- Cost of collection relative to cost of alternative data
- Respondent relationship with BLS
- Concentration of respondents in the sample
- Ease of implementation
- Accuracy issues in the current index...
# In the works

<table>
<thead>
<tr>
<th>Item</th>
<th>RI</th>
<th># quotes</th>
<th>concentration</th>
<th>issues</th>
<th>priority</th>
<th>Source of data</th>
<th>% sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gasoline (all types)</td>
<td>4.344</td>
<td>3,778</td>
<td>M</td>
<td>L</td>
<td>H</td>
<td>scrape</td>
<td>100</td>
</tr>
<tr>
<td>Other motor fuels</td>
<td>0.094</td>
<td>830</td>
<td>M</td>
<td>L</td>
<td>H</td>
<td>scrape</td>
<td>90</td>
</tr>
<tr>
<td>New vehicles</td>
<td>3.695</td>
<td>1,900</td>
<td>L</td>
<td>H</td>
<td>H</td>
<td>sec</td>
<td>100</td>
</tr>
<tr>
<td>Physicians' services</td>
<td>1.728</td>
<td>1,993</td>
<td>L</td>
<td>H</td>
<td>H</td>
<td>sec</td>
<td>75</td>
</tr>
<tr>
<td>Hospital services</td>
<td>2.312</td>
<td>2,123</td>
<td>L</td>
<td>H</td>
<td>H</td>
<td>sec</td>
<td>85</td>
</tr>
<tr>
<td>Cable and satellite television service</td>
<td>1.501</td>
<td>1,906</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>sec</td>
<td>95</td>
</tr>
<tr>
<td>Wireless telephone services</td>
<td>1.693</td>
<td>1,279</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>sec</td>
<td>98</td>
</tr>
<tr>
<td>Land-line telephone services</td>
<td>0.572</td>
<td>874</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>sec</td>
<td>95</td>
</tr>
<tr>
<td>Internet services &amp; electronic info providers</td>
<td>0.780</td>
<td>773</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>sec</td>
<td>95</td>
</tr>
<tr>
<td>Service Type</td>
<td>RI</td>
<td># of Concentration</td>
<td>Priority</td>
<td>Source of Data</td>
<td>Experience</td>
<td>% Sample</td>
<td></td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-----</td>
<td>--------------------</td>
<td>----------</td>
<td>---------------------</td>
<td>------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>Prescription drugs</td>
<td>1.316</td>
<td>4,641</td>
<td>H</td>
<td>H</td>
<td>corp</td>
<td>some</td>
<td></td>
</tr>
<tr>
<td>Limited service meals and snacks</td>
<td>2.542</td>
<td>2,808</td>
<td>M</td>
<td>L</td>
<td>corp</td>
<td>pursue</td>
<td></td>
</tr>
<tr>
<td>Delivery services</td>
<td>0.014</td>
<td>231</td>
<td>H</td>
<td>L</td>
<td>corp</td>
<td>pursue</td>
<td></td>
</tr>
<tr>
<td>Airline fares</td>
<td>0.683</td>
<td>1,745</td>
<td>H</td>
<td>L</td>
<td>scrape, corp</td>
<td>research</td>
<td></td>
</tr>
<tr>
<td>Used cars and trucks</td>
<td>2.329</td>
<td>4,537</td>
<td>H</td>
<td>H</td>
<td>sec Prod, seek</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Postage</td>
<td>0.094</td>
<td>230</td>
<td>H</td>
<td>L</td>
<td>sec prod</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leased cars and trucks</td>
<td>0.655</td>
<td>265</td>
<td>L</td>
<td>H</td>
<td>sec research</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Electricity</td>
<td>2.655</td>
<td>1,406</td>
<td>M</td>
<td>M</td>
<td>seek</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Utility (piped) gas service</td>
<td>0.747</td>
<td>1,404</td>
<td>M</td>
<td>M</td>
<td>seek</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rent and OER</td>
<td>31.548</td>
<td></td>
<td></td>
<td></td>
<td>seek</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- Significant portion of the CPI based on alternative data within 5 years
- Substantial R&D on methodology needed
- Alternative data introduced incrementally alongside monthly publication
Questions for FESAC

- Do you have any reactions to the general criteria CPI has used to date for determining fitness for use? Are we missing anything, etc.?
- Do our criteria for establishing priorities in moving forward make sense to you?
- Any advice for meeting the methodological challenges BLS faces with some of the alternative data sources?
Contact Information

Crystal Konny
Branch Chief
Branch of Consumer Prices
Konny.crystal@bls.gov

Brendan Williams
Senior Economist
Branch of Consumer Prices
Williams.Brendan@bls.gov

David Friedman
Associate Commissioner
Prices and Living Conditions
Friedman.david@bls.gov