On the use of internet data in the CPI

Robert Griffioen and Olav ten Bosch
Overview

– On the use of internet data in the CPI: Challenges of Big Data
  - Population volume
  - Method: price index method and classification method
  - Monthly validation of production process of the CPI
– Big Data
Population volume

- Product type: clothing
- Scale up from 3 to 20 retailers in production (30,000 to 500,000 records)
- Offer Prices

Data collection in stores
Cut-off sample

Internet data
Cut-off complete
Population volume: Big Data?

– No, not from an IT-perspective.
– However, for a Statistical Office they are Big data:
  - Shrinking budget and number of employees
  - Maintaining the quality of an Official Statistics
**Price index method: current situation**

Lowest coicop index

\[ l_{CI}^r_{G,y,m} = \sum_{g \in G} w_{g,m}^r I_{g,y,m}^r \]

Elementary index of a productgroup

\[ I_{g,y,m}^r = \frac{p_{g,y,m}^r}{p_{g,y-1,12}^r} \]

Average of a productgroup

\[ p_{g,y,m}^r = \frac{1}{n_{g,y,m}^r} \sum_{a \in g} \sum_{d \in m} p_{a,y,m,d}^r \]
Classification method: current situation

Text classification

- L-Coicop
- Women
- Men
- Kids
- Clothing type
- Jeans
- Trousers
- Top
- Sweaters
- ...

Classification rule: “sweater” and not “jack”
Classification method: quality

Quality determined by
- Information has to be present
- Classification rules
- Retailers sells goods of the same market segment
- Information quality of the website: is a ‘top’ really a ‘top’?
  → Homogeneity of the groups
Retailers with many different brands from different market segments
– Adapt the current method
– Substitute for a new method
Classification method: future

- More advanced text-mining
- Automatic maintenance of classification rules
  rule: ‘top’ and no (laptop or photoprint)
  → Image processing
- Classification rules replaced by ML-algorithms.

Fine-knit cotton jumper with long sleeves and roll edges around the neckline, cuffs and hem.
Monthly validation process: Currently

<table>
<thead>
<tr>
<th>Role</th>
<th>Robot team</th>
<th>CPI team</th>
<th>Consumption analysis</th>
<th>Calculation &amp; validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>Data validation 1</td>
<td>Data validation 2</td>
<td>Data validation 3</td>
<td>Data validation 4</td>
</tr>
<tr>
<td></td>
<td>Data collection</td>
<td>Checks:</td>
<td>Checks:</td>
<td>Checks:</td>
</tr>
<tr>
<td></td>
<td>Checks:</td>
<td>• Number of items after deduplication</td>
<td>• Changes in groups</td>
<td>• Top down analysis</td>
</tr>
<tr>
<td></td>
<td>• File size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Number of items</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Monthly validation process: Future 1

Role: DevOps team

Process:
- Data validation 1
- Data validation 2
- Data validation 3
- Data validation 4

Data collection:
- Checks:
  - File size
  - Number of items
  - Number of items after deduplication

Consumption analysis:
- Checks:
  - Changes in groups

Calculation & validation:
- Checks:
  - Top down analysis
Monthly validation process: Future 2

Role

DevOps team

Meso-analysis

Process

Data validation 1 → Data validation 2 → Data validation 3 → Data validation 4

Data collection
Checks:
• File size
• Number of items
• Number of items after deduplication

Checks:
• Top down analysis (Downdrill)
• More integrated automatic checks
• Dashboard visualisation
• Weekly indices for validation
• Process information
• Changes in groups
Conclusions

– The CPI production processes are influenced by Big Data:
  - Volume: Big enough for Official Statistics
  - Velocity: higher frequency indices
  - Variety: Different websites with different information and with different information quality → Differences in production processes
– Thank you for your attention!
– Questions?
Indices

031210 Men clothing

031220 Ladies clothing