Implementation of the treatment of the scanner data in France

Guillaume Rateau
Head of the CPI methodology section
Introduction : project schedule

since 2009

Solution studies, implementation of the IT system to treat scanner data and establishing of the legal framework

2018

Double computation of indices without integration of the indices based on scanner data in the CPI

2019

End of the traditional price collection for the scope of the project and replacement by scanner data
Introduction: scope of scanner data

Outlets: supermarkets and hypermarkets
(no discounters, no small-scale retailers)

Products: manufactured food, beverages (01-021),
household goods (0561), pets products (09342),
products for personal care (12132)
(no fresh products: meat, fish, vegetable, fruits)

Geography: mainland France
(no overseas department)

⇒ 14% of the expenditure covered by HICP

Studies: only voluntary retailers data = 30% target expenditure
only 8 consumption segments representative different cases
Introduction : scanner data

Transaction files

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlet</th>
<th>GTIN</th>
<th>Description</th>
<th>Quantity sold</th>
<th>Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>20160608</td>
<td>933</td>
<td>3272770004817</td>
<td>ST MORET PLAIN 150G</td>
<td>1</td>
<td>1.89</td>
</tr>
<tr>
<td>20160608</td>
<td>933</td>
<td>3154230040286</td>
<td>HERTA BACON 150G</td>
<td>2</td>
<td>4.76</td>
</tr>
<tr>
<td>20160610</td>
<td>933</td>
<td>3184670001080</td>
<td>RIANs STRAINED SOFT 6% MG 1KG</td>
<td>1</td>
<td>2.59</td>
</tr>
<tr>
<td>20160610</td>
<td>825</td>
<td>2071900007304</td>
<td>ERSTEIN SUGAR SEMOLINA BEET KG</td>
<td>2</td>
<td>2.70</td>
</tr>
</tbody>
</table>

produced by each outlet

Characteristics files

<table>
<thead>
<tr>
<th>GTIN</th>
<th>Brand</th>
<th>Type of oil</th>
<th>Total volume</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>3265477983004</td>
<td>ISIO 4</td>
<td>MIXTURE</td>
<td>1200 ml</td>
<td></td>
</tr>
<tr>
<td>3760109431149</td>
<td>J LEBLANC</td>
<td>SUNFLOWER</td>
<td>1000 ml</td>
<td>...</td>
</tr>
</tbody>
</table>

≈ twenty characteristics per products, extracted from labels and photos
produced by a market intelligence company

Daily sent data ⇒ each day = 50 million observations, 5GB of raw data
**Introduction : methodology**

**Objective**: « usual » price index concepts

**Producer discounts / relaunches**

\[
\text{COICOP} \supset \text{consumption segment} \supset \text{equivalence class (EQ)} \supset \text{GTIN}
\]

equivalence class

\[
\begin{align*}
\text{GTIN with same characteristics (similar volume, include promotions)} \\
\text{= same product for consumers}
\end{align*}
\]

**Fixed basket** = \{ EQ x outlet \}

**Filters**

- dump filter, outliers in price level
- price changes = outliers / retailer discounts \(\in [-50\%, +100\%]\)
- products sold since more than 30 days
Introduction: methodology

Product aggregation

price [product] = \frac{\text{turnover}}{(\text{quantity sold}) \times (\text{volume of material})}

elementary aggregate = \text{consumption segment} \times \text{outlet}

= \text{geometric Laspeyres prices [1}^{\text{st}}\text{-28}^{\text{th}}\text{] month}

⇔ \text{Substitution of consumer in the same outlet}

higher level indices = \text{usual Laspeyres of elementary aggregates}
Introduction: time aggregation?

daily data $\Rightarrow$ price index based on daily prices?
   goods not bought every day $\Rightarrow$ missing prices?

scanner data price = unit value
CPI price = price offer $\approx$ daily unit value
   product = goods in given outlet, at given day of month
justified approach for goods in supermarkets?
   product = EQ $\times$ outlet $\times$ day of month?
      = EQ $\times$ outlet $\times$ week of month?
      = EQ $\times$ outlet?
$\leftrightarrow$ choice of time aggregation formula

use current quantities
   $\Rightarrow$ differences with unweighted aggregate?
Outline

1. Can daily prices be considered?
2. Time aggregation formula
3. Differences with unweighted aggregates
Outline

1. Can daily prices be considered?
2. Time aggregation formula
3. Differences with unweighted aggregates
1. Daily prices : interpolating

goods not bought every day ⇒ missing prices

several ways of interpolating

Examples:

1. Carry forward
   \[ p_{d+t} = p_d \]

2. Linear interpolating
   \[ p_{d+t} = p_d + \frac{t}{T} (p_{d+T} - p_d) \]

3. Middle point
   \[ p_{d+t} = p_d \text{ if } t < \frac{T}{2}; \quad p_{d+T} \text{ otherwise} \]
1. Daily prices: assessment of the error

1. estimate $E\left[ \frac{P_{d+i}}{P_{d+i}} \mid i \right]$ from data by exhaustive cross-validation
2. compute the expected relative bias for each month

⇒ low level of error
⇒ thereafter, daily prices defined by the middle point method
Outline

1. Can daily prices be considered?
2. Time aggregation formula
3. Differences with unweighted aggregates
2. Time aggregation: formulae

Consider the extreme cases:

- Product = EQ × outlet × day of month
  - Different product each day of the month

\[
\bar{P}^m_i = \prod_{d=1}^{28} (P^{d,m}_i)^{\omega_{i,d}}
\]

\(\omega_{i,d}\) = quantities product i sold during year Y-1 \times price in Dec Y-1

- Product = EQ × outlet
  - Same product during the whole month

\[
\bar{P}^m_i = \frac{\sum_{d=1}^{28} q^{d,m}_i \times P^{d,m}_i}{\sum_{d=1}^{28} q^{d,m}_i}
\]

\(q^{d,m}_i\) = quantity sold day d month m of product i

⇒ different formulae
2. Time aggregation: daily vs monthly prices

Comparison of monthly changes: $I_{v}^{m/m-1} = \prod_{i} \left( \frac{\bar{p}_{i}^{m}}{\bar{p}_{i}^{m-1}} \right)^{\omega_{i}}$

⇒ monthly unit value index more volatile, marked differences
2. Time aggregation: same/different products?

Are the product different in level during the month?

assess day of week effect = mean (residues of moving averages over 7 days)
week of month effect = mean (residues of moving averages over 4 weeks of weekly unit values prices)
2. Time aggregation: same/different products?

⇒ relatively low differences of price levels during the month
2. Time aggregation: same/different products?

Are the paths of prices different during the month?

⇒ monthly changes related to each day of week, each week of month

⇒ very similar price paths
2. Time aggregation: same/different products?

⇒ some paths seem to be different
2. Time aggregation : conclusion

At this stage

Scope : goods (no fresh products) sold in supermarket (2013-2016)

• no structural difference of price levels
• no dynamic difference at the level of the day
  ⇒ no point to consider price index based on daily prices
  
• dynamic differences at the level of the week
  ⇒ are they due to discounts?
Outline

1. Can daily prices be considered?
2. Time aggregation formula
3. Differences with unweighted aggregates
3. Differences: discounts / relaunches

2 cases

- producer discounts / relaunches
  - change packaging,
  - extra volume offer,
  - relaunches, ...

  ⇒ different barcodes

  treated through the equivalence classes

- retailer discounts
  - reduced price,
  - extra product offer
  - ...

  ⇒ same barcodes

  treated when computing the price by the unit value

  Hereafter, retailer discounts = drop of more than 20% price
3. Differences: producer discounts / relaunches

Computation of the monthly changes without the equivalence classes

⇒ differences between indices are not due to producer discounts/relaunches
3. Differences : retailer discounts

Computation of the monthly changes without the retailer discounted products

⇒ differences between indices are mainly due to retailer discounts
3. Differences : retailer discounts

What are these discounts? ⇒ focus on the olive oil

- Small sales share (\(\approx 2.5\%\)), very short duration (\(\leq 4\) days in average) generally related to an increase of quantities ... but not always
- ... and also explosion of quantities
- Tiny part of very high discounts (up to 90%) are they outliers?
Conclusion

For the scope of goods (no fresh products) sold in supermarket (2013-2016)

• no structural differences of prices within the month
  ⇒ no need to define a daily or weekly prices index
  ⇒ prices can be computed as a monthly unit values

• may exist marked differences between price indices using fixed weights and current quantities
  ⇒ differences are mainly due to very short & important retailer discounts
  ⇒ compared to “traditional” CPI, change of weights put on discounts
  ⇒ fine tuning of the price change filter ?
Thank you for your attention