Scanner data and multilateral price index methods

UN GWG on Big Data for Official Statistics
Workshop on Scanner Data and Official Statistics
Kigali, Rwanda, 29 April – 1 May 2019

Delivering insight through data for a better Canada
Outline

• Three classes of multilateral index methods:
  • GEKS method
  • Geary-Khamis method
  • Time Product Dummy (or fixed effects) method

• Length of index estimation window
• Extension methods
• Monitoring and quality control
• Conclusion
Classes of multilateral index methods

• Methods traditionally used in the spatial comparison of price levels
• Comparison should be independent of the choice of base country/region:
  • methods are transitive
• Recently adapted to temporal comparison of price levels
• Use weights at product level
• Are free of chain drift
• Multilateral methods and their applications:

<table>
<thead>
<tr>
<th>Multilateral method</th>
<th>Expenditure data needed?</th>
<th>Known application</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEKS-Törnqvist</td>
<td>Yes</td>
<td>Australia: Grocery data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>New Zealand: Electronics</td>
</tr>
<tr>
<td>GEKS-Jevons</td>
<td>No</td>
<td>Does not take advantage of available sales data!</td>
</tr>
<tr>
<td>Geary-Khamis (GK)</td>
<td>Yes</td>
<td>Netherlands: Almost all scanner data including supermarkets data</td>
</tr>
<tr>
<td>Time Product Dummy (TPD)</td>
<td>Yes/No, both versions possible</td>
<td>New Zealand: Rental prices</td>
</tr>
</tbody>
</table>
GEKS method

• Gini (1931); Eltető and Köves (1964) and Szulc (1964)
• Estimation window $[0, T]$. For $t$ in $[0, T]$:

$$
P^{0,t} = \prod_{z=0}^{T} \left( \frac{P^{0,z}}{P^{t,z}} \right)^{\frac{1}{T+1}}
$$

• $T+1$ bilateral price indexes $P^{0,z}, z = 0, 1, \ldots, T$
• $T+1$ bilateral price indexes $P^{t,z}, z = 0, 1, \ldots, T$
  • $P^{0,t}$ is calculated through $T+1$ paths: $P^{0,z}P^{z,t}$
  • None of these paths should be preferred to the others
  • Geometric average of these $T+1$ results
GEKS method

- Bilateral index formula:
  - Needs to satisfy the time reversal test
  - Fisher index
  - Törnqvist index
  - Jevons index (not so relevant in the context of scanner data!)

- If the dataset has limited/no characteristics but we have expenditure information, we can use a superlative bilateral index formula (e.g. Törnqvist, Fisher)

- If the dataset has no expenditure information, we can use an unweighted bilateral formula (e.g. Jevons)
Geary-Khamis method

- Geary (1958) and Khamis (1972)
- For an homogeneous product, in a given time period, price equals unit value:
  - Unit value = \( \frac{\text{total sales value}}{\text{total quantity sold}} \)
  - \( p^t_i q^t_i = \left( \frac{p^t_i}{v_i} \right) \left( v_i q^t_i \right) \) for a product \( i \) available in period \( t \)
- \( v_i \) is a quality adjustment factor for product \( i \).
- \( \left( \frac{p^t_i}{v_i} \right) \) : Quality adjusted prices; \( \left( v_i q^t_i \right) \) : Adjusted quantities
- Adding up quantities of dissimilar goods to form the unit value index isn’t necessarily meaningful
- Use standardized or quality-adjusted quantities
- Apply quality adjustment factors to the various item quantities to express them in terms of a “base” product, and then simply add them up
Geary-Khamis method

• Quality adjusted unit value:

\[
p^t = \frac{\sum_{i \in G^t} \left( \frac{p_i^t}{v_i^t} \right) (v_i q_i^t)}{\sum_{i \in G^t} v_i q_i^t} = \frac{\sum_{i \in G^t} p_i^t q_i^t}{\sum_{i \in G^t} v_i q_i^t}
\]

“Quality adjusted prices” and “adjusted quantities” world

• GK price index between 0 and \( t \):

\[
p^{0,t} = \frac{p_t}{p^0}
\]

\[
p^{0,t} = \frac{\left( \sum_{i \in G^t} p_i^t q_i^t \right) / \left( \sum_{i \in G^t} v_i q_i^t \right)}{\left( \sum_{i \in G^0} p_i^0 q_i^0 \right) / \left( \sum_{i \in G^0} v_i q_i^0 \right)}
\]

(1) “Change in total sales from 0 to \( t \)”, “Weighted quantity index”

(2) “Weighted deflated prices”: reference prices

• Need to solve (1) and (2) simultaneously; iterative algorithms typically used
Time Product Dummy method

- Country Product Dummy method in the spatial price comparison
- Product prices in month $t$ follows a stochastic model:

$$\ln p_i^t = \alpha + \delta^t + \gamma_i + \epsilon_i^t$$

- Estimation method:
  - Weighted least squares regression
  - Expenditure shares $s_i^t$ as weights

- Let $v_i = \exp\left(\hat{\gamma}_i\right)$. It can be shown that for a set of products $G^t$,

$$\overline{p}^t = \prod_{i \in G^t} \left(\frac{p_i^t}{v_i}\right)^{s_i^t}$$

Quality adjusted price
Time Product Dummy method

• Time Product Dummy index between 0 and \( t \):

\[
\frac{P^{0,t}}{P^0} = \frac{\prod_{i \in G^t} \left( \frac{p^t_i}{v_i} \right)^{v_i}}{\prod_{i \in G^0} \left( \frac{p^0_i}{v_i} \right)^{v_i}}
\]

Ratio of quality adjusted prices

\[
v_i = \prod_{z=0}^{t} \left( \frac{p_i^z}{P_i^{0,z}} \right)^{w_i^z} ; \quad w_i^z = \frac{s_i^z}{\sum_{k=0}^{t} s_i^k}
\]

Quality adjustment factors

• Close similarities with GK method
• Indexes can be calculated using the regression framework or an iterative algorithm as for GK
Length of index estimation window

• Estimation windows shorter than a year:
  • Problematic with seasonal products

• Estimation windows larger than a year:
  • Differences with respect to 13-month window estimation are generally small for published classes level
  • Bilateral indexes between all pairs of months of the window are used:
    • A very large window may lead to a loss of characteristicity
    • The estimated price change does not actually pertains only to the two periods under comparison

• A 13-month window is typically used
  • ABS chose a window of 5 quarters for their quarterly CPI
Extension methods

- Why index series extension?
- Index estimation based on a fixed window \([0, T]\)
- Data from new period \(T+1\) can alter comparisons between earlier periods
- CPI is non revised
- How do we form a multilateral “window” incorporating the current period?
- How do we splice the results onto previous index levels?
Extension methods

Three period TPD

<table>
<thead>
<tr>
<th>Fixed Effect Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

\[ I_{TPD}^{AB} = \frac{\exp(\delta^B)}{\exp(\delta^A)} = 1.35 \]

Four period TPD

<table>
<thead>
<tr>
<th>Fixed Effect Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

\[ I_{TPD}^{AB} = \frac{\exp(\delta^B)}{\exp(\delta^A)} = 1.45 \]
Extension methods

• Rolling or expanding window approaches

Rolling window

<table>
<thead>
<tr>
<th>time</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>t</th>
<th>t+1</th>
<th>t+2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Expanding window

<table>
<thead>
<tr>
<th>time</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>t-1</th>
<th>t</th>
<th>t+1</th>
<th>t+2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Fixed length
• Variable start point

• Variable length
• Fixed start point (can be updated from time to time)
Extension methods

• How could we link together indexes from successive windows?

• Link in one period (which one?)

• Take the geomean over all links
Monitoring and quality control

- Important to have clearly defined checks and analyses on the data:
  - So large that any manual verification will not work

- Quality control before index calculation:
  - Has each new file been read correctly?
  - Plot time series of sample sizes
  - Plot time series of the number of unique products classified to each CPI lowest level class
  - Plot time series of the total sales of all products classified to each CPI lowest level class
Monitoring and quality control

• Quality control after index calculation
  
  • Decide on which elementary aggregates/published classes need to be reviewed after index calculation
  
  • Develop tools to identify CPI aggregates with unusual month over month price changes or unusual twelve-month price changes
    
    • Investigate on the main drivers of the identified ‘outliers’ and decide on their treatment
  
  • Develop decomposition tools that help with explaining price movements
Conclusion

• Multilateral index methods:
  • use sales and quantity data
  • give transitive index formula
  • are free of chain drift

• Product definition should be determined before using a multilateral index method

• All three method classes generally give similar price index results, but not exactly the same index values!
Conclusion

• Development of multilateral index methods has benefited from the availability of big and rich scanner data

• Do participants countries’ statistical agencies already have access to retail scanner data?

  • How important are the consumer purchases made from retailers equipped with scanner registers in the different countries?

  • Are retailers not willing to cooperate in sharing their scanner data?

  • Any other issues?
Questions?

Thank you!