Multilateral indices and the relaunch problem

Product clustering and alternative solutions

Jacco Daalmans
8 June 2022
Contents

- Relaunch problem
- Product clustering
- Alternative solutions: imputation and matching
- Conclusions
Relaunch problem
Relaunches

- Product replaced by (almost) identical one
- Price change often larger than quality change
- Matched-model methods miss price changes due to relaunches (if nothing is done)
- Here focus on: Geary-Khamis, TPD and GEKS-Törnqvist
Product clustering
Method

- Items combined into product clusters
- Cluster prices computed as unit values
- Unit value bias for heterogeneous items
- Crucial: Suitable cluster definition
Example: Unit value bias

<table>
<thead>
<tr>
<th></th>
<th>Price (0)</th>
<th>Price (1)</th>
<th>Quantity (0)</th>
<th>Quantity (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product 1</td>
<td>10</td>
<td>12</td>
<td>100</td>
<td>300</td>
</tr>
<tr>
<td>Product 2</td>
<td>18</td>
<td>20</td>
<td>300</td>
<td>100</td>
</tr>
<tr>
<td>Unit value price</td>
<td>16</td>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Clustering: pros and cons

Pros
- Broad applicability
- Well established
- Easy

Cons
- Unit value bias (heterogeneous strata)
- Analysis below cluster level impossible (loss of details)
Imputation
Method

- Price estimated for non-sold items

- Aim: solve relaunch problem rather than to complete data -> might give interpretation problems e.g. imputed price for a nonseasonal item

- Non-trivial choice between imputation methods

- Methods available for some indices (e.g. GEKS-Törnqvist) but less well known for others (e.g. TPD)
Relation imputation and clustering

Imputation methods exist that give the same results as product clustering

- GEKS- Törnqvist:
  clustering same as imputing each price with unit value
  (unobserved and observed prices)

- TPD and Geary-Khamis:
  Same as for GEKS- Törnqvist, but also quantities need to be imputed
Relation imputation and clustering

- Unit value bias in clustering corresponds to a rigorous imputation approach

- In paper: new imputation methods that:
  - mimic product clustering
  - less rigorous replacement of prices

If all prices are observed -> No adjustment (contrary to clustering)
If many prices are missing -> Similar results as clustering
Product matching
Method

- Each new product matched with a disappeared product
- Replacement and replaced products should be similar
- Semi-automatic procedures proposed: text mining and manual analysis
- Automatable methods needed for transaction data
- Solution needed if number of new and disappeared products are unequal
- In paper: a simple procedure with arbitrary choices
Matching

Pros
No adjustment of prices and quantities (contrary to clustering and imputation)

Cons
Arbitrary choices (matching procedure)

Interpretation of the matched products
Simulation
Setup

- 12 month data TV’s, chocolate and potato products
- Fixed population: products that have been sold each month
- Simulated relaunches (change of product ID)
- Monte carlo simulation (100 replicates)
- Comparison:
  - no correction, clustering, imputation, matching versus ‘true’ index without simulated relaunches
- Criterion: median abs difference of index values
Scenario 1: random relaunches

<table>
<thead>
<tr>
<th></th>
<th>TV’s: Median distance from true index</th>
<th>Chocolates: Median distance from true index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GEKS-Törnvist</td>
<td>Geary Khamis</td>
</tr>
<tr>
<td>No correction</td>
<td>1.18</td>
<td>1.32</td>
</tr>
<tr>
<td>Imputation</td>
<td>0.65</td>
<td>4.74</td>
</tr>
<tr>
<td>Matching</td>
<td>1.38</td>
<td>1.48</td>
</tr>
<tr>
<td>Clustering</td>
<td>4.56</td>
<td>5.52</td>
</tr>
</tbody>
</table>
Scenario 1: random relaunches

<table>
<thead>
<tr>
<th>Potatoes:</th>
<th>GEKS-Törnqvist</th>
<th>Geary Khamis</th>
<th>TPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>No correction</td>
<td>1.96</td>
<td>1.40</td>
<td>1.22</td>
</tr>
<tr>
<td>Imputation</td>
<td>1.78</td>
<td>1.31</td>
<td>1.15</td>
</tr>
<tr>
<td>Matching</td>
<td>0.48</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td>Clustering</td>
<td>1.27</td>
<td>1.37</td>
<td>1.37</td>
</tr>
</tbody>
</table>

- Best correction method:
  - For TV’s and Chocolates: ‘No correction’
  - For Potatoes: ‘Matching’

- Matching always better than clustering
Scenario 2: Non-random relaunches

- Relaunches occur at one time period for randomly selected 75% of all items.

- Each relaunch goes along with a simulated, permanent price increase by 20%.
## Scenario 2: Non-random relaunches

### TV’s: Median distance from true index

<table>
<thead>
<tr>
<th>Method</th>
<th>GEKS-Törnqvist</th>
<th>Geary Khamis</th>
<th>TPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>No correction</td>
<td>11.14</td>
<td>7.39</td>
<td>10.54</td>
</tr>
<tr>
<td>Imputation</td>
<td>1.24</td>
<td>3.81</td>
<td>5.11</td>
</tr>
<tr>
<td>Matching</td>
<td>2.10</td>
<td>1.44</td>
<td>3.62</td>
</tr>
<tr>
<td>Clustering</td>
<td>5.63</td>
<td>6.66</td>
<td>6.70</td>
</tr>
</tbody>
</table>

### Chocolates: Median distance from true index

<table>
<thead>
<tr>
<th>Method</th>
<th>GEKS-Törnqvist</th>
<th>Geary Khamis</th>
<th>TPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>No correction</td>
<td>14.77</td>
<td>11.90</td>
<td>14.61</td>
</tr>
<tr>
<td>Imputation</td>
<td>1.83</td>
<td>0.34</td>
<td>3.04</td>
</tr>
<tr>
<td>Matching</td>
<td>0.20</td>
<td>4.09</td>
<td>0.54</td>
</tr>
<tr>
<td>Clustering</td>
<td>0.74</td>
<td>0.85</td>
<td>0.81</td>
</tr>
</tbody>
</table>
# Scenario 2: Non-random relaunches

<table>
<thead>
<tr>
<th>Potatoes:</th>
<th>Median distance from true index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GEKS-Törnqvist</td>
</tr>
<tr>
<td>No correction</td>
<td>13.33</td>
</tr>
<tr>
<td>Imputation</td>
<td>0.83</td>
</tr>
<tr>
<td>Matching</td>
<td>0.95</td>
</tr>
<tr>
<td>Clustering</td>
<td>1.10</td>
</tr>
</tbody>
</table>

- Correction method necessary ("no correction" gives large errors)

- Best correction method depends on data set and index method:
  * Matching (5 cases)
  * Imputation (3 cases)
  * Clustering (1 case)

- Matching better than clustering for 7 out of 9 cases
Conclusion
- Correction for relaunches needed (given a price increase)

- Clustering easy, well-understood, broad applicable, but unit value bias for heterogeneous strata

- Imputation and matching mostly give better results in a simulation study (especially matching)

- Drawbacks: arbitrary choices, unnatural to ‘add’ values to transaction data, interpretation difficulties.
Thank you!