Session 2.1

- The use of scanner data in Denmark - a case study
Data are validated in 3 steps

- First step: Simple checksystems in Sas (missing records, proper dates, number of records and format)
- Second step: SD is linked to the COICOP-classification based on the supermarkets’ own classification and a searchword-process created in Excel and linked to Sas
- Third step: Further cleaning and aggregation of scanner data
Step 1: What is the output of validation-system

- Check for missing records
- Check for proper dates
- Check for number of records
- Check for format
EXERCISE 1 (5 minutes)

• What kind of check systems would you like to have for scanner data?
Step 2: Linking of Scanner data to ECOICOP

• The linking of GTIN codes to COICOP is done at a 6 digit level for each supermarket type of store. At the moment we use existing 8 digit level where possible
How is the linking done?

- First we have the nomenclature, which can be updated:

<table>
<thead>
<tr>
<th>C6</th>
<th>C6_description</th>
<th>C6_description_extra</th>
</tr>
</thead>
<tbody>
<tr>
<td>11110</td>
<td>Rice</td>
<td>Rice, porridge of rice, ricedesserts</td>
</tr>
<tr>
<td>11121</td>
<td>Flower and grain</td>
<td>Wheat flower, wheat grains, other types of flower</td>
</tr>
<tr>
<td>11131</td>
<td>Ryebread</td>
<td>All kind of rye bread</td>
</tr>
<tr>
<td>11199</td>
<td>Residual group of bread and bakery products</td>
<td></td>
</tr>
</tbody>
</table>
• Then we have the nomenclature linking (looked through once a year)

<table>
<thead>
<tr>
<th>xxx_VNR</th>
<th>xxx_C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxxxx</td>
<td>11110</td>
</tr>
<tr>
<td>xxxxxx</td>
<td>11199</td>
</tr>
</tbody>
</table>
What do we do with the residual groups

<table>
<thead>
<tr>
<th>C6</th>
<th>C6_description</th>
<th>Searchword1</th>
<th>Searchword2</th>
<th>New_C6</th>
<th>New_C6_description</th>
</tr>
</thead>
<tbody>
<tr>
<td>11199</td>
<td>Residual group of bread and bakery products</td>
<td>%BASMATI%</td>
<td></td>
<td>11110</td>
<td>Rice</td>
</tr>
<tr>
<td>11199</td>
<td>Residual group of bread and bakery products</td>
<td>%HVEDEMEL%</td>
<td></td>
<td>11121</td>
<td>Flower and grain</td>
</tr>
</tbody>
</table>
How does this look in real life?
How do we monitor the new search words needed

<table>
<thead>
<tr>
<th>C6</th>
<th>C6_description</th>
<th>eggp Turner</th>
<th>turnover_c2_a_ggr</th>
<th>turnovershare of C2-level (PCT)</th>
<th>turnover_c4_a_ggr</th>
<th>turnovershare of C4-level (PCT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2299</td>
<td>Restgrupper tobak</td>
<td>105.110</td>
<td>165.113</td>
<td>98.79</td>
<td>165.113</td>
<td>98.79</td>
</tr>
<tr>
<td>1339</td>
<td>Restgrupper fisk</td>
<td>3.202.726</td>
<td>1.447.052.237</td>
<td>0.22</td>
<td>52.538.203</td>
<td>6.1</td>
</tr>
<tr>
<td>11198</td>
<td>Restgrupper af bageri</td>
<td>8.491.246</td>
<td>1.447.052.237</td>
<td>0.58</td>
<td>184.288.644</td>
<td>4.61</td>
</tr>
<tr>
<td>11299</td>
<td>Restgrupper kød og fjørkø</td>
<td>10.576.361</td>
<td>1.447.052.237</td>
<td>0.73</td>
<td>269.208.928</td>
<td>3.93</td>
</tr>
</tbody>
</table>
What isn’t linked then?

<table>
<thead>
<tr>
<th>Eancode</th>
<th>Product description</th>
<th>C6</th>
<th>C6 description</th>
<th>Ean-tornaover</th>
<th>The eancodes’ share of the residual group</th>
<th>The whole residual groups’ share of the C2-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>123456789101LAKS</td>
<td>11399 Restgrupper fisk</td>
<td>183.9</td>
<td>4.95</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>123456789101MOERKSEFILET</td>
<td>11399 Restgrupper fisk</td>
<td>299.3</td>
<td>8.06</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### And what do we do

- We use

<table>
<thead>
<tr>
<th>C6</th>
<th>C6_description</th>
<th>Searchword1</th>
<th>Searchword2</th>
<th>New_C6</th>
<th>New_C6_description</th>
</tr>
</thead>
<tbody>
<tr>
<td>11399</td>
<td>Residual group of fish</td>
<td>%MOERK5E</td>
<td>FILET%</td>
<td>11311</td>
<td>Cod and the like</td>
</tr>
</tbody>
</table>
Step 3: Aggregation

• This includes making amounts uniform
• It also includes cleaning other variables (product text, unit, amount…) at the lowest possible level

• For example GTINs with more than one amount are recalculated into numeraire
• Per default all PLUs with more than one product text, amount or unit are deleted here!
• It also includes removing prices, which are more than a factor 4 away from the median price for a given GTIN

• Now: The prices are aggregated into one price per GTIN per type of shop
• **The final step:**
  • More weeks are aggregated together (including cleaning) to form the data for one month

• **Application of filters:**
  • Prices which have changed by more than a factor 10, since last month, are removed
  • Prices which have fallen in both price and volume since last month are removed
CPI via scanner data

- Reception of scanner data
  - Data is validated
    - Data is ready for processing
      - The representative basket (a sample) for month t-1 is drawn
        - Each good in the representative basket is checked - is it still sold? yes/no?
          - Yes
            - The representative basket for month t is drawn
          - No
            - Replacement goods are chosen
Selection criteria (1)

- Goods in the representative basket need to have been sold for all 12 months of 2011 (2015) and constitute the top 50 % of turnover.

- There are individual criteria at either too many or too few observations compared to the existing sample.
Selection criteria (2)

- Replacement goods need to have been part of the sample for 4 months
- They are then picked based on highest turnover and stability of turnover
System for maintenance of the representative basket

![Excel spreadsheet with labels]

- Column A: Year (4 cifre)
- Column B: Month (2 cifre)
- Column C: Week (2 cifre)

1. Dan data
2. Hent udgåede varer
5. Opdater kurv
6. Dan data til Oracle
On the calculation of SD-indices

- Calculation is based on Laspeyres type indices (JEVONS)- done at all levels following the existing calculation methods

- February 2015: Commenced with the calculation of RYGEKS indices, so we are able to have a basis for a juxtaposition.

→ suggest small underestimation of price development, but rygeks method is not perfect
Documentation?
COICOP 1
COICOP 2
No significant differences
Scanner data is a better source
Scanner data is a different source
Worries

- The complexity of the data can involve the risk of a suboptimal sample
- Data is not delivered timely
Emergency systems

• We use two weeks for the calculation of SD-indices giving us a chance of switching weeks.

• Developed SAS-programs for the continuation of prices for one or more SD-chains

• We’ve signed contracts with the chains on the delivery of data. This will include a clause letting us know if the product structure changes
Lessons learnt

1. Scanner data is complex. Cleaning of data and an assumption of diversity of barcodes is necessary
2. The differences when handling scanner data as compared to traditional price collection need to be implemented in the IT-systems in a proper way and thought through as early as possible
3. Common frame work is important
On unit values

• In general unit values are applicable for the calculation of SD-indices

• But:
  - The data needs to be cleaned in other variables before hand (product text, unit, amount…) at the lowest possible level
  - The price development has to be monitored both within the month and between months
End discussion

• What are your worries with the processing of scanner data?