Selective editing and automated correction of micro data in the SBS, Norway

Jakob Kalko, Statistics Norway
Agenda

• Background

• Editing of enterprises – R-type consistency indicator

• Editing of MEE’s (multiple establishment enterprises) - P-type consistency indicator

• Automated correction of SBS data
Background

• For years, large focus at micro level. Flagging of soft/absolute errors. Too much editing at microlevel?

• Publishing of current SBS: t +17 months.

Several reasons for improving this:
- relevance of the data in general
- give NA access to data earlier
- future demands from Eurostat concerning earlier delivery of data?
- increased demand to effectiveness within SSB
Enterprises. R-type consistency indicators

• *Statistical unit is enterprise*. Unweighted indicators at micro level are established, showing the development in the relationship between two selected variables in year t and t-1

• Weighted indicators at micro level are established, giving higher weights to large units (measured by employment or man-years)

• Indicators are established at macro level (2-5 digit NACE), based on indicators at micro level

• We chose to focus on three indicators and one total indicator, based on the three indicators.
R-type consistency indicator - micro level

• Target variable: Employment  Auxiliary variable: Man-years.

• Un-weighted R-type consistency indicator \((R_{em})\):  
  
  \[ R_{em} = \frac{\text{employment t} / \text{man-years t}}{\text{employment t-1} / \text{man-years t-1}} \]

• Weighted R-type consistency indicator \((R_e)\)

  Weight \((w)\):  
  
  \[ \sqrt{\text{Employment t} / \sqrt{10}} \]

• \(R_e = R_{em}^w\)
R-type consistency indicator - micro level

- **Target variable: Wages**  
  **Auxiliary variable: Man-years.**

- **Un-weighted R-type consistency indicator** \((R_{vm})\):
  \[ R_{vm} = \frac{\text{wages}_t / \text{man-years}_t}{\text{wages}_{t-1} / \text{man-years}_{t-1}} \]

- **Weighted R-type consistency indicator** \((R_{a})\)
  
  Weight \((w)\):
  \[ \sqrt{\text{man-years}_t} / \sqrt{10} \]

- \[ R_{a} = R_{vm}^w \]
R-type consistency indicator - micro level

• Target variable: Turnover  Auxiliary variable: Employment

• Un-weighted R-type consistency indicator \( (R_{te}) \):
  \[
  R_{te} = \frac{\text{turnover } t}{\text{employment } t} / \frac{\text{turnover } t-1}{\text{employment } t-1}
  \]

• Weighted R-type consistency indicator \( (R_t) \)
  
  Weight \((w)\): \(\sqrt{\text{Employment } t / \sqrt{10}}\)

• \( R_t = R_{te}^w \)
Total weighted R-type consistency indicator

- Total weighted R-type consistency indicator \((R_t^W)\)

\[ R_t^W = (R_e \times R_a \times R_t)^{1/3} \]

- At micro level a R-indicator=1, indicates no change in the relationship between the target variable and auxiliary variable.
R-type consistency indicator, macro level

- Given a R-type consistency indicator $R$ for the e-th unit, in the n-th NACE

- Let $s_{ne} =$ turnover for the e-th unit, in the n-th NACE

- Let $S_{ne} =$ turnover for the population for the units e=1,2...E, in the n-th NACE

- $W_e = s_{ne} / S_{ne}$

- **R-type consistency indicator, macro level** $(R_m)$ for $R_t$:
  
  $R_m = \sum_{e=1}^{E} W_e \ log_{10} (R_t)$. 

- $R_m = 0$ indicates no change at macro level
Practical use of R-type consistency indicator

• Preliminary data 2011
  - 3. digit nace groups with $R > 0.1$ were listed out
  - Enterprises within these nace groups were listed out, including the R-indicator.
  - Units with unreasonable indicators were examined/corrected
    New lists were created. Editing stopped when indicator became stable
  - The R-indicators \((\text{turnover/employment})\) and the total weighted indicator showed up to be especially useful
P-type consistency indicator

• Is made for MEE’s and is measuring the development in the relationship between two selected variables in year t and t-1

• Takes into account the weights of the variables measured in the MEE in the reference periods t and t-1 (Fischer index)

• Practical use: No practical experience but intention is to follow the same principles as for the R-indicator:
  - MEE’s where P-indicator differs significantly from 1 are localized. Establishments being the worst outliers in the MEE are listed and examined. New P-indicator is then created.
P-type consistency indicator – practical example

<table>
<thead>
<tr>
<th>Establishment</th>
<th>Turnover (t)</th>
<th>Operation. costs (t)</th>
<th>Turnover (t-1)</th>
<th>Operation. costs (t-1)</th>
<th>R - indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>K=1</td>
<td>18</td>
<td>15</td>
<td>20</td>
<td>12</td>
<td>0,7200</td>
</tr>
<tr>
<td>K=2</td>
<td>12</td>
<td>11</td>
<td>14</td>
<td>12</td>
<td>0,9350</td>
</tr>
<tr>
<td>K=3</td>
<td>35</td>
<td>28</td>
<td>34</td>
<td>35</td>
<td>1,2868</td>
</tr>
<tr>
<td>P_{Dir}^{t-1}</td>
<td>65</td>
<td>54</td>
<td>68</td>
<td>59</td>
<td>1,0444</td>
</tr>
</tbody>
</table>

L-index: \((20/68) \times 0,72 + (14/68) \times 0,935 + (34/68) \times 1,2868=1,0477\)

P-index: \((18/65) \times 0,72 + (12/65) \times 0,935 + (35/65) \times 1,2868=1,0649\)

F-index: \(\sqrt{1,0477} \times \sqrt{1,0649} = 1,0563\)

P-indicator = F-index / (P_{Dir}^{t-1}) = 1,0563/1,0444=1,0114
Practical experiences – R and P-indicators

• The R-indicator $R_t$ \((\text{turnover/employment})\) and the total weighted indicator $R_{tW}$ showed up to be especially useful.

• Indicators can not replace the administration of the population.

• Reasonable indicators are not necessarily an indication of no significant changes in absolute variables (eg. mergers/splits).

• Unreasonable indicators might be correct (weak connection between variables creating the indicator, mergers/splits).
Practical experiences – R and P-indicators

• Ir-regular R-indicators in MEE’s should be treated differently in the estimation of the P-indicator.

• No significant experience, using the P-indicator so far. Should consider further whether the current indicator gives us the information we need.
Automatic correction of SBS data

- Basic idea: Replace manual corrections of certain variables in MEE’s with automatic corrections, to increase productivity.

- Following variables are received broken down from enterprise level to establishment level through the SBS survey:
  
  *Employment, wages, turnover, operational costs and gross investments.*

- Only wages, turnover and operational costs are automatic corrected.

- Variables are only corrected if they do not sum up to enterprise level.
Automatic correction of SBS data

• New distribution of data among establishments:
  - using keys based on the raw data or
  - using keys based on an alternative distribution based on year t-1

• Wages are corrected, based on man-years

• Turnover is corrected, based on employment

• Operational costs are corrected, based on turnover
## Automatic corrections – practical example

<table>
<thead>
<tr>
<th>Period</th>
<th>t</th>
<th>t-1</th>
<th>Turnover, raw data</th>
<th>Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Establishments</td>
<td>Employment</td>
<td>Employment</td>
<td>t</td>
<td>t-1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4.468</td>
<td>7.561</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2.431</td>
<td>840</td>
</tr>
<tr>
<td>Sum, establish.</td>
<td>4</td>
<td>5</td>
<td>6.899</td>
<td>8.401</td>
</tr>
<tr>
<td>Sum, enterprise</td>
<td>4</td>
<td>5</td>
<td>8.985</td>
<td>8.401</td>
</tr>
<tr>
<td>Difference</td>
<td>0</td>
<td>0</td>
<td>-2.086</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Establishments</th>
<th>Alternative suggested distribution, Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: (7.561/4) x 2</td>
<td>3.780,5</td>
</tr>
<tr>
<td>2: (840/1) x 2</td>
<td>1.680,0</td>
</tr>
<tr>
<td>Sum</td>
<td>5.460,5</td>
</tr>
<tr>
<td>Difference</td>
<td>5.460 – 8.985 = -3.524,5</td>
</tr>
</tbody>
</table>
**Automatic corrections – practical example**

- Distribution of turnover in this example will be based on the distribution of raw data.
- Sum of turnover should = 8.985 (enterprise level)

<table>
<thead>
<tr>
<th>Establishment</th>
<th>Turnover, raw-data</th>
<th>Distribution</th>
<th>Corrected turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>4.468</td>
<td>0.648</td>
<td>5.819</td>
</tr>
<tr>
<td>2</td>
<td>2.431</td>
<td>0.352</td>
<td>3.166</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>6.899</strong></td>
<td><strong>1.000</strong></td>
<td><strong>8.985</strong></td>
</tr>
</tbody>
</table>
Automatic corrections - experiences

• Increased effectiveness, editing large MEE’s

• In MEE’s where data for only one establishment is given, no automatic corrections are done for the other establishments

• MEE’s in industries with weak connection between turnover and employment should be paid extra attention

• Establishments founded in year t-1 might have a different relationship between variables in year t. May lead to errors in the automatic correction

• In the long run – automatic corrections may be based on earlier automatic corrections. Problem? Solutions?