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 choosed 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} = (employment t / man-years t) / (employment t-1 / man-years t-1)

- **Weighted R-type consistency indicator ($\mathbf{R}_{\boldsymbol{e}}$)**

  Weight (w): $\sqrt{Employment\ t}$ / $\sqrt{10}$

- $\mathbf{R}_{\boldsymbol{e}}$**=** $\mathrm{R}_{\mathrm{em}}{}^{\mathrm{w}}$

R-type consistency indicator - micro level

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

- **Un-weighted R-type consistency indicator (R_{vm}):**

  R_{vm} = (wages t /man-years t) / (wages t-1 / man-years t-1)

- **Weighted R-type consistency indicator ($\mathbf{R_a}$)**

  Weight (w): $\sqrt{man - years\ t}$ / $\sqrt{10}$

- $\mathbf{R_a}$= $\mathrm{R_{vm}}^{\mathrm{w}}$

R-type consistency indicator - micro level

- *Target variable: Turnover* *Auxiliary variable: Employment*

- **Un-weighted R-type consistency indicator (R_{te}):**

  R_{te} = (turnover t / employment t) / (turnover t-1 / employment t-1)

- **Weighted R-type consistency indicator ($\mathbf{R_t}$)**

  Weight (w): $\sqrt{Employment\ t}\ /\ \sqrt{10}$

- $\mathbf{R_t}$= $\mathrm{R_{te}}^{\mathrm{w}}$

Total weighted R-type consistency indicator

- **Total weighted R-type consistency indicator ($\boldsymbol{R_{tW}}$)**

- R_{tW} = $(\mathbf{R}_{\boldsymbol{e}} \text{ x } \mathbf{R}_{\boldsymbol{a}} \text{ x } \mathbf{R}_{\boldsymbol{t}})^{1/3}$

- 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 nacegroups 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 (*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 referenceperiods 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

Establish-ment	Turnover (t)	Operation. costs (t)	Turnover (t-1)	Operation. costs (t-1)	R - indicator
K=1	18	15	20	12	0,7200
K=2	12	11	14	12	0,9350
K=3	35	28	34	35	1,2868
$P_{Dir}^{t-1,t}$	**65**	**54**	**68**	**59**	**1,0444**

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

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

F-index: $\sqrt{1{,}0477}\ x\ \sqrt{1{,}0649}$ = 1,0563

P-indicator =F-index /**(**$P_{Dir}^{t-1,t}$) = 1,0563/1,0444=*1,0114*

Practical experiences – R and P-indicators

- The R-indicator $\mathbf{R}_{\mathbf{t}}$ *(turnover/employment*) and the total weighted indicator $\boldsymbol{R}_{\boldsymbol{tW}}$ 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 wether 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

Period	t	t-1	t	t-1
Establishments	Employment	Employment	Turnover, raw data	Turnover
1	2	4	4.468	7.561
2	2	1	2.431	840
Sum, establish.	**4**	**5**	**6.899**	**8.401**
Sum, enterprise	**4**	**5**	**8.985**	**8.401**
Difference	*0*	*0*	*- 2.086*	*0*

Establishments	Alternative suggested distributrion, Turnover
1: (7.561/4) x 2	3.780,5
2: (840/1) x 2	1.680,0
Sum	**5.460,5**
Difference	*5.460 – 8.985 = -3.524,5*

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)

Establishment	Turnover, raw- data	Distribution	Corrected turnover
1:	4.468	0,648	5.819
2	2.431	0,352	3.166
Sum	*6.899*	*1,000*	*8.985*

Automatic corrections - experiences

- Increased effectivenes, 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 ?