28thVoorburg Group Meeting

Tokyo, Japan

7- 11 October 2013

Selective editing of data and automated corrections of micro-data in the Structural Business Statistics in Norway

Jakob Kalko

Jakob.Kalko@ssb.no

+47 62 88 54 95

Selective editing of data and automated corrections of micro-data in the Structural Business Statistics

1. Summary

Statistics Norway introduced for the reference year 2011 two types of statistical indicators, in order to to increase effectiveness, editing the sample in the Structural Business Statistics (SBS). In addition automatic micro processing of SBS data were implemented.
Statistical indicators for enterprises, R-type consistency indicator, increased effectiveness of editing of enterprise data. This effect was especially significant in connection with preliminary data. Statistical indicators for establishment, P-type consistency indicator, were not used to a large extend for the reference year 2011. This indicator might also need to be supplemented with other indicators at establishment level. Automatic corrections of employment, turnover, wages and operational costs contributed to increased effectiveness of editing, but have also weaknesses, which the editor should be aware about.

2. Background

For many years Statistics Norway (SSB) has produced Structural Business Statistics. The statistic is a sample survey and traditionally the total sample has been edited, to create the best possible basis for estimation of data for the total population. The micro-perspective has been dominating in this process. The editing has been done on the basis of several controls (soft and absolute) for each unit. Controls have been (and are still) present at both establishment and enterprise level. In order to increase effectiveness and being able to publish the final data earlier than 15 months after the end of the reference period SSB took the following steps for the reference period 2011:

- Reduction of the sample, mainly reduction of the number of small units.
- Introducing editing after statistical indicators. The purpose is to localize the significant units in the sample and put the attention to these. The purpose is also to reduce the total scope of editing by putting more focus at the macro perspective
- Automatic corrections of data at establishment level in enterprises with more than one establishment. The purpose here is also to reduce the scope of editing the process.

The paper will not discuss the reduction of the sample.

Li-Chun Zhang from the Division of statistical methods in Statistics Norway has been contributing significantly during the process, with advices and written input. A large part of the paper is based on his input during the process, nevertheless any mistakes in the paper are purely the responsibility of the author.

3. R-type consistency indicators

3.1 R-type consistency indicator – micro level

Denote by (u,v) that values of which the ratio u/v is an indicator of interest. Denote by (u t-1, v t-1) the corresponding values at t-1. The unweighed R-type consistency indicator at micro level is then as follows:

Table 3.1. Calculation of unweighed R-type consistency indicator, micro level

(u,v)	**(>0,>0)**		**(0/-, 0/-)**		(0/-,>0	(>0,0/-)
(u t-1, v t-1)	(>0,>0)	Incomplete	(>0,>0)	Incomplete	-	-
R	(u/v)/(u t-1/v t-1)	-1	-2	-3	-8	-9

Source: Li-Chun Zhang, SSB.

Indicators where the formula (u/v)/(u t-1/v t-1) can be used are mentioned as *regular indicators.*
Indicators with values between -1 and -9 are called *irregular indicators*.
The rest of the description is focusing on regular indicators.

SSB decided to focus on 4 R-type indicators at micro level, which are defined in the boxes below. It should be added that it is of course possible to create other indicators than the ones mentioned below.

a). 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 ($\mathbf{R}_{e}$)

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

$\mathbf{R}_{e} = R_{em}^{\ w}$

b). 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 ($\mathbf{R}_{\square}$)

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

$\mathbf{R}_{\square} = R_{vm}^{\ w}$

c) 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): $\sqrt{employment\ t}\ /\ \sqrt{10}$

$\mathbf{R}_{t} = R_{te}^{\ w}$

A total weighted R-type consistency indicator ($\mathbf{R}_{tw}$) can be calculated by:

$\mathbf{R}_{tw} = (\mathbf{R}_{e} \times \mathbf{R}_{\square} \times \mathbf{R}_{t})^{1/3}$

At micro level a R-indicator=1 indicates no changes in the relationship between the target variable and the auxiliary variable. Concerning the weighted R-indicators, 10 (man-years or employment) is used as the balance point between small units (low importance) and large units (high importance) based on the structure of most industries. It could be argued to increase it to 20 or perhaps higher. One could also argue that this number should differ from industry to industry. Anyway, in this first year of practicing the R-indicator, we have chosen to have one common "balance point" for all industries . It is undecided whether this will be changed for the reference year 2012.

3.2 R-consistency indicator – macro level.

Based on the indicators at micro level, a R-indicator is created at macro level. By macro level is meant 2-5 digit nace level.

Box 3.1 Calculation of R-type consistency indicator, macro level.

Given a R-type consistency indicator R_t 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 perfect match at macro level.

3.3 Practical use of the R-type consistency indicator

Macro indicators are used to examine whether the development between two (or more, e.g $\mathbf{R}_{tw}$**)** variables from t-1 to t are "reasonable" at a given nace-level. If indicators differ too much from zero, micro indicators are listed, to find the enterprises which contribute most to this development. There is no statistical rule for when the R-indicators are defined as unreasonable. Indicators are listed out at different nacelevels. Within business services (section L, M, N and S) attention were put towards macro indicators (nace groups) $R > 0.1$. Units were listed out, including the R-indicator for each of them. Unreasonable indicators were examined, or either corrected or documented. The experience was positive, especially using the indicators R_t and R_{tw} in connection with the publishing of the preliminary data. It became easier to quickly localize unreasonable variables. In appendix I, the table shows the development of the indicators during the editing process. R-Indicators for each enterprise are also visible in the application for editing the SBS.

4. P-type consistency indicators

4.1 Calculation of P-type consistency indicator

The indicator is only made within multi establishment enterprises (MEE). It shows the aggregated development for establishments in an MEE between two variables from t-1 to t.
In order to take into account the different weights of the target variable in t-1 and t, the estimation includes a Laspeyre index (t-1), and Paasche index (t) which gives the possibility of estimating a Fischer index. The mathematical formula for creating the P-type indicator is as follows:

Box 4.1: Mathematical definition of the P-type indicator

1 Given a R-type consistency indicator for a group of units, k=1,……K

$R_k^{t-1,t} = (u_k^t / v_k^t) / (u_k^{t-1} / v_k^{t-1})$

2. A Laspeyre index with a price index-like construction for the whole group is given with:

$$P_L^{t-1,t} = \sum_{k=1}^{K} w_k^{t-1} R_k^{t-1,t} = \sum_{k=1}^{K} (u_k^{t-1} / \sum_{j=1}^{K} u_j^{t-1}) R_k^{t-1,t}$$

3. A similar constructed Paasche index is like this:

$$P_P^{t-1,t} = \sum_{k=1}^{K} w_k^{t} R_k^{t-1,t} = \sum_{k=1}^{K} (u_k^{t} / \sum_{j=1}^{K} u_j^{t}) R_k^{t-1,t}$$

4. The Fisher index is then given by:

$P_F^{t-1,t} = (P_L^{t-1,t} \times P_P^{t-1,t})^{1/2}$

5. Direct R-type consistency indicator for the whole group:

$P_{Dir}^{t-1,t} = [(\sum_{k=1}^{K} u_k^t) / (\sum_{k=1}^{K} v_k^t)] / [(\sum_{k=1}^{K} u_k^{t-1}) / (\sum_{k=1}^{K} v_k^{t-1})]$

6. A P-type consistency indicator is given by:

$P^{t-1,t} = P_F^{t-1,t} / P_{Dir}^{t-1,t}$

Source: Li-Chun Zhang, SSB.

Table 4.1 Practical example. NOK million. Estimation of the P-type indicator

Establishment	Turnover (t)	Op. costs (t)	Turnover (t-1)	Op. 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
Enterprise ($P_{Dir}^{t-1,t}$)	**65**	**54**	**68**	**59**	**1,0444**
				L-index	1,0477
				P-index	1,0649
				F-index	1,0563
				P- type indicator	**1,0114**

L-index ($P_L^{t-1,t}$): (20/68) × 0,7200 + (14/68) ×0,9350 + (34/68) ×1,2868= 1,0477

P-index ($P_P^{t-1,t}$): (18/65) × 0,7200 + (12/65) × 0,9350 + (35/65) ×1,2868 =1,0649

F-index ($P_F^{t-1,t}$)= $\sqrt{1,0477} \times \sqrt{1,0649}$ =1,0563

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

4.2 Practical use of the P-type consistency indicator

The P-type consistency indicator were not used to a large extend for the reference year 2011, but some general thoughts concerning how it should be used has been made..

Enterprises with P-type consistency indicator values varying significantly from 1 are localized. There is no statistical definition of extreme values. As for the R-indicators, a list is made and based on this the worst outliers are localized. We can use the R-type indicator for each establishment, to find the

units, contributing most to an unreasonable P-type-indicator for the enterprise. P-type indicators are also directly visible in the application for editing the SBS.

For both P- and R-consistency indicators, it applies that they are used to direct the editor in the right direction of the editing process. When no significant change in the indicator occurs between two periods of editing, the process should be stopped.

5. Some general limitations using the indicators.

- Using R-type and P-type indicators can not replace the administration of the population (evaluation of NACE code at micro-level, change of activity etc.)
- A P-type indicator close to 1 may be the result of a "unstable" development among establishments. We should establish supplementing indicators for MEE's. One opportunity could be measuring the standard deviation among the R-indicators in the MEE.
- Mergers/splits may not cause significant changes in the indicators, but might influence the absolute figures significantly
- Within industries with weak connection between e.g. turnover/employment and turnover/wages large changes in indicators might be correct and may cause unnecessary extra work
- In MEE's where one establishment has an ir-regular R-indicator, the total P-indicator is also defined as ir-regular. Since we are losing some important information according to this definition this is to be changed for the reference year 2012, creating the P-indicator only based on establishments with regular R-indicators.

6. Automatic micro processing of SBS data

6.1 Description of the process

Within MEE's SSB collect information concerning employment, wages, turnover, operational costs and gross investments for each establishment. In the application for editing these data from year t-1 is available, useable for controlling the data from. Several numbers of MEEs consist of a large number of establishments. Editing these takes a lot of resources. From the reference year 2011 we therefore introduced automatic micro processing of these data. The processing has some basic rules:

- Only wages, turnover and operational costs are corrected
- Variables are only corrected if the do not sum up to enterprise level (based on the account)
- Corrections are based on information based on rawdata year t or data from year t-1
- Wages are corrected based on man-years
- Turnover is corrected based on employment
- Operational costs are corrected based on turnover

Where a correction is needed, an alternative distribution of the variable based on information from year t-1 is estimated.. The sum of the raw data and the alternative distribution are compared to the sum at enterprise level. The distribution which sum up closest to the enterprise level is used. Data are then corrected according to 1) this distribution and 2) to the sum at enterprise level.

Table 6.1 Practical example. Automatic correction of turnover (NOK 1000)

Period	t	t-1	t	t-1	t
Establishments	**Employm.**	**Employm.**	**Turnover raw data**	**Turnover**	**Corr. turnover**
1	2	4	4 468	7 561	5 819
2	2	1	2 431	840	3 166
Sum, establishment	**4**	**5**	**6 899**	**8 401**	**8 985**
Sum enterprise	**4**	**5**	**8 985**	**8 401**	**8 985**
Difference	0	0	-2 086	0	0

Table 6.2 Suggested alternative distribution of turnover, with employment as the auxiliary variable

Establishments	**Turnover**
1. (turnover t-1/employment t-1) × employment t	3780,5
2. (turnover t-1/employment t-1) × employment t	1680,0
SUM	**5460,5**

*The suggested alternative distribution **is not used** since the sum of the raw data (6 899) is closer to the sum at enterprise level (8 985). Keys based on the distribution of raw data are used to correct the sum and distribution of turnover. Corrected data can be viewed in the column **Corr. Turnover in table 6.1**.*

Turnover is corrected before the costs are corrected. The practical example in table 6.3 takes the results in table 6.1 into account.

Table 6.3 Practical example. Automatic correction of operational costs (NOK 1000)

Period	t	t-1	t	t-1	t
Establishments	**Corr. Turnover**	**Turnover**	**Op. Costs, rawdata**	**Op.costs**	**Corr. Op. costs**
1	5 819	7 561	3 340	7 158	5 152
2	3 166	840	1 944	795	2 802
Sum, establishments	**8 985**	**8 401**	**5 284**	**7 953**	**7 954**
Sum enterprise	**8 985**	**8 401**	**7 954**	**7 953**	**7 954**
Difference	0	0	-2 670	0	0

Table 6.4 Suggested alternative distribution of operational costs, with turnover as the auxiliary variable

Establishments	**Op. costs**
(costs t-1/turnover t-1) × turnover t	5 509
(costs t-1/turnover t-1) × turnover t	2 996
SUM	**8 505**

*The suggested alternative distribution is used since the sum (8 505) is closer to the sum at enterprise level (7 954) than the sum of the raw data (5 284). Keys based on the alternative distribution are used to correct the sum and distribution of costs. Corrected data can viewed in column **Corr. Op. costs.***

A practical example where variable t-1 = 0 can be viewed in appendix II

6.2 Experiences and challenges using automatic corrections

- Significant increased effectiveness, editing large MEE's
- Some cases where raw-data were closer to the truth than the corrected data occurred e.g. in establishments where employment was stable, but turnover had changed a lot.
- MEE's where only data for one establishment were given. In some cases this equal to the enterprise data = no corrections were made.
- Special attention should be put to industries where connection between turnover and employment are weak.
- Over time – be aware that if only automatic corrected data are used, future corrections will also be based on these – and not on the information given by the respondent. In the long run, this is a factor of uncertainty which should checked.

7. Closing remarks

In the future SSB will try make even better use of the statistical indicators and automatic corrections in the SBS. It might be that the P-consistency indicator will be supplemented with other statistical analysis. One option to be considered is to measure the standard deviation of the R-indicators of the establishments in each MME.

It should also be considered whether the ideas implemented for the SBS are transferrable to other statistics within SSB. The methodology will also be considered as a possible source for increased functionality in ISEE, which is SSB's common framework application for editing and estimation of data.

Appendix I

Development in the total weighted R-type consistency indicator R_{tw}

Nace	1'st run	2'nd run	3'rd run	Last run	Difference two last runs	Difference 1st and last run
68.1	0,09	0,04	0,04	0,04	0,00	-0,05
68.2	0,13	0,04	0,04	0,04	0,00	-0,09
68.3	0,15	0,08	0,08	0,08	0,00	-0,07
69.1	0,40	0,08	0,08	0,05	-0,02	-0,34
69.2	0,36	0,12	0,12	0,12	0,00	-0,24
70.2	0,21	0,08	0,05	0,05	0,00	-0,16
71.1	0,37	0,15	0,15	0,15	0,00	-0,22
71.2	0,39	0,48	0,48	0,48	0,00	0,09
72.1	0,34	0,42	0,41	0,41	0,00	0,07
72.2	0,23	0,24	0,24	0,24	0,00	0,01
73.1	0,11	0,06	0,06	0,06	0,00	-0,05
73.2	0,43	0,05	0,05	0,05	0,00	-0,38
74.1	0,17	0,06	0,06	0,06	0,00	-0,11
74.2	0,04	0,03	0,03	0,03	0,00	0,00
74.3	0,08	0,06	0,06	0,06	0,00	-0,03
74.9	0,10	0,07	0,06	0,06	0,00	-0,04
75.0	0,03	0,03	0,03	0,03	0,00	-0,01
77.1	0,25	0,06	0,06	0,06	0,00	-0,19
77.2	0,15	0,06	0,06	0,06	0,00	-0,09
77.3	0,16	0,11	0,11	0,11	0,00	-0,05
77.4	1,12	0,25	0,25	0,24	-0,01	-0,87
78.1	0,13	0,08	0,08	0,08	0,00	-0,06
78.2	0,94	0,30	0,30	0,30	0,00	-0,64
79.1	0,19	0,14	0,07	0,07	0,00	-0,12
79.9	0,21	0,07	0,07	0,07	0,00	-0,14
80.1	0,58	0,26	0,25	0,25	0,00	-0,34
80.2	0,17	0,09	0,09	0,09	0,00	-0,08
80.3	0,06	0,04	0,04	0,04	0,00	-0,01
81.1	0,32	0,08	0,08	0,08	0,00	-0,24
81.2	0,44	0,11	0,11	0,11	0,00	-0,33
81.3	0,02	0,01	0,01	0,01	0,00	-0,01
82.1	0,15	0,05	0,05	0,05	0,00	-0,11
82.2	0,40	0,14	0,14	0,12	-0,02	-0,28
82.3	0,19	0,07	0,07	0,07	0,00	-0,13
82.9	0,15	0,06	0,06	0,06	0,00	-0,09
95.1	0,17	0,06	0,06	0,06	0,00	-0,11
95.2	0,07	0,03	0,03	0,03	0,00	-0,04
96.0	0,15	0,05	0,05	0,05	0,00	-0,11

Appendix II

Practical example target variable > 0 year t, target variable=0 year t-1

Periode	t	t-1	t	t-1	t	
Establishments	Employm.	Employm.	Turnover	Turnover	Corr. turnover	Estimation
1	15	0	10 000	0	16 406	(15/32) x 35000
2	17	0	20 000	0	18 594	(17/32) x 35000
Sum, establishments	**32**	**0**	**30 000**	**0**	**35 000**	
Sum, enterprise	**32**	**0**	**35 000**	**0**	**35 000**	
Difference	**0**	**0**	**-5 000**	**0**	**0**	

References

Li-Chun Zhang: Kort om editering av strukturstatistikk, foreløpige tall. 2012. Internal document

Li-Chun Zhang: Automated micro-processing of SBS data. 2012. Internal document