Implementation of NACE Rev. 2 in sample surveys

Milda Šličkutė-Šeštokienė

Statistics Lithuania, Lithuania
e-mail: milda.slickute-sestokiene@stat.gov.lt

Abstract

The paper describes the process of implementation of NACE Rev. 2 in enterprise surveys. Particular attention is paid for estimation of data of Quarterly Survey on Earnings by NACE Rev. 2 when reference period is before 2008. The problems analyzed and solutions accepted are presented.

1 Introduction

Implementation of NACE Rev. 2 is one of the main tasks in all EU institutions of national statistics as well as in Statistics Lithuania. It is extremely time-consuming work witch requires good understanding of statistical process.

Among other tasks, related to implementation of NACE Rev. 2, estimation of back data is one of the most complicated. From one point of view Statistics Lithuania has the obligation to inform policy makers and other users with high quality data but from another point of view back data estimation is extremely costly procedure and that means that careful decision has to be made on back data estimation policy. All necessary back data have to be estimated using reasonable and limited resources.

The implementation strategy for Quarterly Survey on Earnings is presented in this paper. The results of this survey are one of the first data that will be published by NACE Rev. 2. The first publication of data of Quarterly Survey on Earnings by NACE Rev. 2 is foreseen in May 2009 for reference period – first quarter of 2009.

Short plan for implementation of NACE Rev. 2 for Quarterly Survey on Earnings:

- Since 2009 data are published only by NACE Rev. 2;
- For 2008 – data are presented by both NACE (at the stage of sample selection it is foreseen to get data by both NACE);
- 2005-2007 – back data estimation applying micro-approach;
- 2000-2004 – back data estimation applying macro-approach.

2 Sample selection by both NACE

Quarterly Survey on Earnings is a sample survey. Stratified simple random sample is used. Stratification criterions are as follows: by NACE Rev 1.1 (usually at two digit level, sometimes more detailed), by economic sector (public or private) and by size of enterprise (1-9, 10-49, 50-99, 100-249, 250-499, 500 and > employees). Horvitz-Thompson estimator is used for estimation.

For reference year 2008 at the moment of sample selection it was foreseen to get results by both NACE. Three different approaches of sample selection were analyzed:

- To select a sample only by NACE Rev. 1.1, but significantly enlarge the sample size so that reliable estimates by NACE Rev. 2 could be achieved;
- To select two independent samples by two different NACE;
- To select sample by NACE Rev. 1.1 and later to select some additional enterprises in those economic activities by NACE Rev. 2 where sample size is not sufficient.

Sample only by NACE Rev. 1.1 could give low precision for results by NACE Rev. 2 especially in new NACE sections. Also two different samples would guarantee better precision but it would be too big burden for respondents and for staff of Statistics Lithuania. So it was decided to use third option which is compromise between those two mentioned above and also compromise between quality, burden and cost of the survey. Final sample size is 8014 (16.2% of total) enterprises or 685310 (54.7% of total) employees. Reference year 2008 is the only year when sample is drawn by both NACE.

3 Micro approach versus macro approach

Micro approach versus macro approach was analyzed for estimation of back data of Quarterly Survey on Earnings for period 2000-2007. It was decided to combine those two approaches in order to get most efficient one in the terms of cost and data quality.

3.1 Micro approach (reference period 2005-2007)

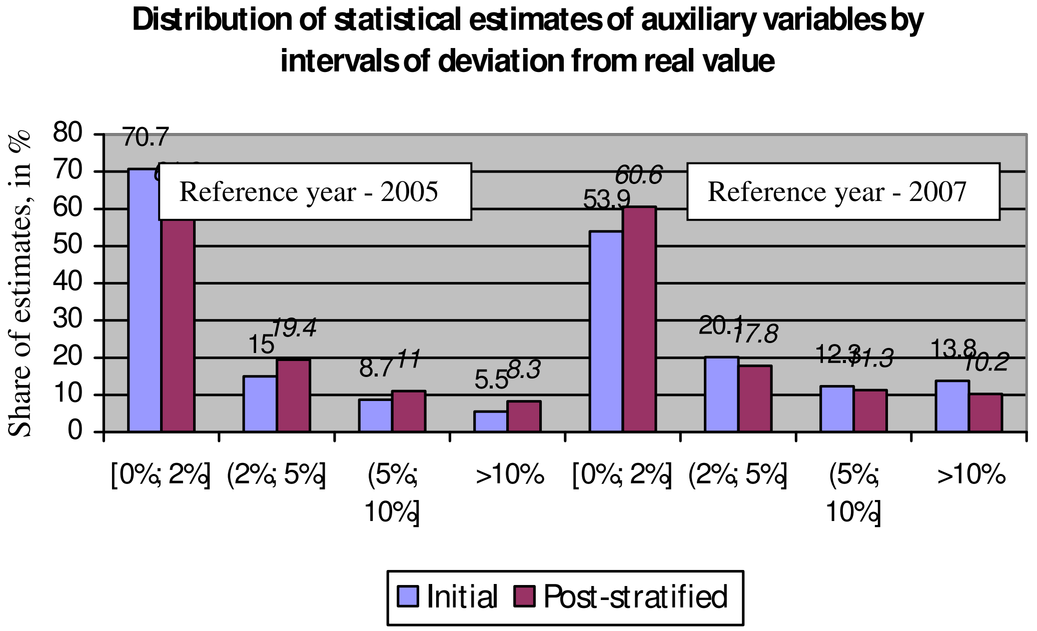
Lithuanian Register of Statistical Entities is double coded by NACE Rev. 1.1 and NACE Rev. 2 for reference period 2005-2007. Micro approach will be used for back data estimation for this period. Two main tasks have to be solved for estimation by NACE Rev. 2 using micro approach:

- Which weights will be used: Initial sample weights or post-stratified weights by NACE Rev. 2 (post-stratified weights are calculated as follows: new strata by NACE Rev. 2 are constructed, sample size and population size is calculated in each new stratum. Population size is divided from sample size in each new stratum in order to get post-stratified weights).

- What kind of estimator will be used: Since reliable auxiliary information is available it is supposed to use not direct estimator but other estimator based on auxiliary information.

Initial sample weights were compared with post-stratified weights in order to choose which ones are most suitable for estimation by NACE Rev. 2. For that purpose auxiliary variables (data of Social Insurance) were used. The auxiliary variables are strongly correlated with variables of interest (more than 0.9 almost in each domain of interest) and also values of auxiliary variables are known for each element of the frame. It is supposed that error of estimates for variables of interest is about the same as for auxiliary variables. In order to find out which weights are most suitable for estimation by NACE Rev. 2, estimates for auxiliary variables were compared with real values for reference year 2005 and 2007 using different weights. In the picture 1 it is presented the distribution of statistical estimates for auxiliary variables by intervals of deviation from real value. From the picture 1 one could notice that for the reference year 2005 estimates by NACE Rev. 2 with initial weights produce better precision of the results: 70.7 per cent of statistical estimates based on initial weights deviate from real value up to 2 percent; corresponding number for post-stratified weights is 61.3 percent. The data precision for 2005 is higher when initial weights but not post-stratified are used. Analyzing corresponding figures for reference year 2007, the situation is vice versa: using initial weights 53.9 per cent of statistical estimates deviate from real value up to 2 percent and using post-stratified weights 60.6 per cent of statistical estimates deviate from real value up to 2 per cent. So for the reference year 2007 the data precision is higher when post-stratified weights are used.

Picture 1



When estimating data by NACE Rev. 2 depending on economic activity and on reference period, sometimes initial weights produce better quality of the results and sometimes post-stratified weights. Generally there is no significant difference which weights to use. As auxiliary variables are strongly correlated with variables of interest the same conclusion could be made for the variables of interest. It was decided that for estimation of statistical variables by NACE Rev. 2 initial sample weights will be used in order to maintain better comparability between NACE Rev. 1.1 and NACE Rev. 2.

As at the moment of sample selection it was not possible to foreseen to get results by NACE Rev.2 for the reference periods before 2008, for that reason it could happen that quality of the results will be not sufficient by NACE Rev. 2. In order to improve the quality of the results, ratio estimator was analyzed using data of Social Insurance as auxiliary information. Horvitz-Thompson estimator was compared with ratio estimator for key variables. The distribution of coefficients of variation for key variables of interest for Horvitz-Thompson and ratio estimators by NACE Rev. 2 for each quarter of reference years 2007 and 2005 are presented below.

Table 1

Distribution of coefficients of variation of key variables "Number of employees" and "Gross earnings" for Horvitz-Thompson (HT) and ratio (R) estimators by NACE Rev. 2, in per cent

CV	Number of employees								Gross earnings							
	2007 Q1		2007 Q2		2007 Q3		2007 Q4		2007 Q1		2007 Q2		2007 Q3		2007 Q4	
	R	HT	R	HT	R	HT	R	HT	R	HT	R	HT	R	HT	R	HT
[0; 3]	97.3	71.6	97.3	72.3	97.2	69.0	96.6	67.6	98.6	67.6	98.6	68.2	97.9	64.8	97.9	62.8
(3; 5]	1.4	16.2	1.4	13.5	0.7	14.5	1.4	15.9	.	14.2	.	10.8	0.7	11.0	0.7	17.2
(5; 10]	0.7	8.8	0.7	9.5	1.4	9.0	1.4	8.3	0.7	15.5	0.7	16.2	0.7	17.2	0.7	12.4
(10; 30]	0.7	3.4	0.7	4.7	0.7	6.9	0.7	7.6	0.7	2.7	0.7	4.7	0.7	6.2	0.7	6.9
(30; 100]						0.7		0.7						0.7		0.7
Median	0.4	1.5	0.4	1.6	0.5	1.7	0.4	1.8	0.2	1.6	0.2	1.6	0.2	1.7	0.2	1.7

CV	Number of employees								Gross earnings							
	2005 Q1		2005 Q2		2005 Q3		2005 Q4		2005 Q1		2005 Q2		2005 Q3		2005 Q4	
	R	HT	R	HT	R	HT	R	HT	R	HT	R	HT	R	HT	R	HT
[0; 3]	92.1	80.1	90.1	80.8	90.7	79.3	84.7	78.0	98.7	72.8	99.3	72.8	99.3	73.3	98.7	68.0
(3; 5]	6.6	12.6	9.3	13.2	7.3	15.3	13.3	16.7	0.7	15.9	.	17.2	.	18.0	0.7	23.3
(5; 10]	1.3	6.0	.	4.6	1.3	3.3	2.0	3.3	.	9.9	.	7.9	0.7	6.7	0.7	7.3
(10; 30]	.	0.7	0.7	0.7	0.7	1.3	.	2.0	0.7	0.7	0.7	1.3	.	1.3	.	1.3
(30; 100]	.	0.7	.	0.7	.	0.7				0.7	.	0.7	.	0.7		
Median	0.6	1.0	0.7	1.1	0.8	1.2	0.9	1.3	0.1	1.0	0.2	1.1	0.1	1.1	0.1	1.2

The coefficients of variation for ratio estimator are smaller for each variable and for each reference period compared to Horvitz-Thompson estimator. For that reason for back data estimation it was decided to use ratio estimator.

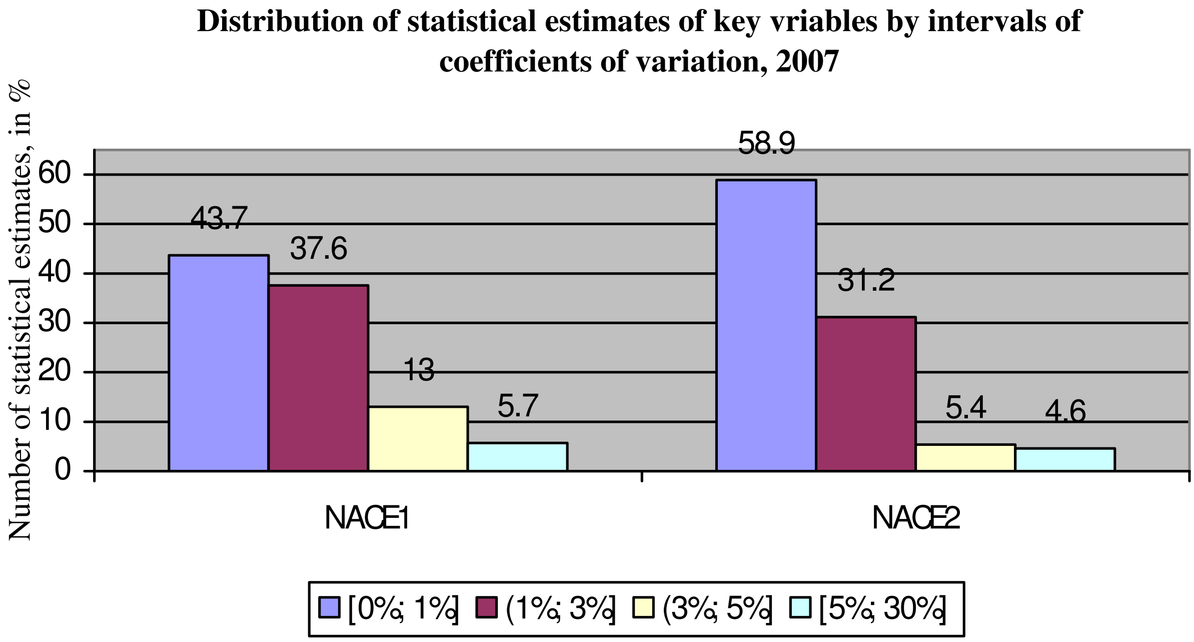
3.2 Macro approach (reference period 2000-2004)

Back data for reference year 2000-2004 also have to be estimated but Lithuanian Register of Statistical Entities will not be double coded for this period. It is supposed to apply macro-approach in order to get data for this period.

3.3 Results achieved

At this moment preliminary data by NACE Rev. 2 are estimated only for reference years 2005 and 2007. The coefficients of variation of the results by NACE Rev. 2 were compared with coefficients of variation of the results by NACE Rev. 1.1. In the picture 2 it is presented the distribution of statistical estimates for key variables by intervals of coefficients of variation for reference year 2007, similar situation was found for the reference year 2005. As could be noticed from picture 2 the quality of the results by NACE Rev. 2 are even better than by NACE Rev. 1.1, 43.7% of coefficients of variation for estimates by NACE Rev 1.1 are up to 1, while for estimates by NACE Rev. 2 the share of coefficients of variation up to 1 is equal to 58.9%. That happened because of ratio estimator (which was applied for estimation by NACE Rev. 2) usually has smaller coefficients of variation compare to Horvitz-Thompson estimator (which was applied for estimation by NACE Rev. 1.1).

Picture 2



Distribution of statistical estimates of key vriables by intervals of coefficients of variation, 2007

References

Sarndal C. E., Swensson B.,Wretman J. (1992): Model Assisted Survey Sampling. Springer-Verlag, New York.

Sarndal C. E., Lundstrom S. (2005): Estimation in Surveys with Nonresponse. John Wiley & Sons, Ltd.

Wallgren A. and Wallgren B. (2006): Register-Based Statistics – Administrative Data for Statistical Purposes. John Wiley & Sons, Ltd.

Eurostat (2007): Back casting handbook.