Frame Error Impact on Structural Business Statistics Surveys

Hysni Elshani

Kosovo Agency of Statistics
e-mail: hysni.elshani@rks-gov.net
hysni.el@gmail.com

Background

This paper treats some details of the first data analysis of the Structural Business Statistics (SBS) concerning the frame error non-response rates and post-stratification as well as the analysis made in order to check and edit data for KAS. The analysis is done on the 'raw' data with higher overview on the frame error, editing, post-stratification, non-response analysis and the short overview on some findings on estimated numbers of enterprises, entrepreneurs and employees in Kosovo.

1. Introduction

The SBS is based on the use of the Statistical Business Register (SBR) as the reference frame. It represents the population of interest for the sampling of units and for the grossing up of sampled data. Quality of statistics produced by each survey is then related to the quality of the SBR. The level of errors in the register and the errors in the sampled-based estimation are then correlated.

The accuracy of estimates depends on their variability and bias. Their magnitude determines the overall error.

This paper aims to identify and measure the impact of principal frame errors on the sample-based estimations using auxiliary administrative variables.

2. Auxiliary administrative information for turnover: the fiscal turnover

The use of administrative sources for statistical purposes continues to be one of the strategic purposes of any statistical institutes. But the possibility to substitute direct information with available administrative data is dependant, where the needed information exists, on their quality *(data from Tax authority of Kosovo).* On a yearly basis, businesses that are liable for VAT are obliged to present the VAT declaration at the Tax Authority offices. In our country, according to the law, subjects that must present yearly VAT declaration are whoever carries out an economic activity (any form of enterprise.

The fiscal turnover figure is going to be used as quantitative variable only in the last years thanks to an improvement of coverage and timeliness from the fiscal administration. Comparisons lead to some inconsistencies due to different reasons, first of all the lack of quality in the BR administrative variable (mainly a certain amount of outliers and missing data). But correlation between fiscal turnover and SBS turnover is very high in fact what a business declares to fiscal authorities is the same it declares in a statistical questionnaire. For this reason, before considering the possibility to substitute fiscal turnover to the surveyed turnover, we analyse results using it as an auxiliary variable in estimation, *(Fiscal turnover is refering the same reference period as the SBS survey data refer).*

3. The frame errors implications on sampled - based estimation

It is known that the purpose of each survey is to produce estimate as accurate as possible of a given unknown parameter. Sampling and non-sampling errors determine the level of quality of sample-based estimates in fact they cause bias and a loss of efficiency. Among non-sampling errors non-responses and coverage problems in the frame of reference represent the main sources of error.

These two factors are correlated because some non-responses can be attributed to errors in the frame such as the impossibility to contact the unit included into the target population as well as incorrect information in the frame determines the necessity to delete some unit in the sample reducing its size.

Frame errors and their impact of the overall error have been classified according to the following types:

a) under - coverage - BR does not reflect businesses within scope for that survey. Reasons for under-coverage errors are well known: omission (lags and leakage), errors in the determination of the state of activity of units (falsely not active units), and mistakes in stratification variables (out of scope units when they are in scope). BR under-coverage generally affects estimations increasing bias;

b) over-coverage - BR considers in scope businesses that are not. Reasons for over-coverage are the opposite of the under-coverage ones: duplication, errors in the determination of the state of activity of units (falsely active units), and mistakes in stratification variables (in of scope units when they are not in scope). Over-coverage generally affects estimations increasing their bias; moreover if a sampled unit is correctly identified as ceased, a reduction of the sample size determines an increase in the sampling error, *in these case we exclude those enterprise from frame*. A specific attention has to be given to errors due to incorrect information held by units correctly registered. Coding errors typically affect stratification, variables such as principal economic activity codes, size in terms of employment, location variables and demographic data, *In case of wrong activity code we contacted recalling businesses in order to ensure the right activity code*. With regards to errors in the BR location variables when a .nit is localised in a different place, *here is mentioned to address or location of surveyed business*. This unit, apart from the fact that it is ceased or active, is located in some other place however this result is treated as a non-response. *It's treated as nonresponse if this unit wasn't part of a sample.* The impact of this error is both on bias (a respondent unit will represent the missing one but it can significantly be different) and sampling variance (reduction of the sample size. For each contacted unit (a response, both in presence of a well filled questionnaire and a blank one) it is possible to obtain information about the correctness of frame variables. In this way some over-coverage problems or inaccurate information can be detected and can give an overall idea of their extent to the whole frame. Errors are classified and grouped together in order to measure their impact on estimations.

4. Some findings regarding the SBS survey in KAS

- *Sampling and survey coverage* - the general rule is to cover at least 80 percent of activity, notably 80 percent of turnover from business register. Several levels were chosen for stratification: (i) first stratification level - by activity by NACE four digits (small activities were sometimes combined in one group); (ii) second stratification level - by size (initially three strata of size class by activity), which in standard SBS is measured based on the number of employees in the unit, but in our case is measured by size of turnover; (iii) third stratification level was within 4 digit by size within the third class. For each stratum initial sample level is defined (mainly 80 percent coverage, plus targeted confidence interval (e.g., the expected rate of non-response).

Thus, we used stratified random sample techniques. The resulted sample size is 3151 enterprises.

The main data source for the Business register is Ministry of Trade and Industry and Tax Authority of Kosovo the information are updated in quarterly basis.

The sample of the survey was designed in 2 fazes:

- one part exhaustive for all enterprises with turnover more than 50 000 euros
- sample for the enterprises with the turnover less than 50000 euros, detailed for each activity at 2 digits level of NACE classification which have more 10 worker.

In 2014, the frame of the survey was 36880 units, from which have been taken for sample 3151, which represent 8.54 % of all active enterprises. It’s very important to say that all these enterprises which have been selected for survey should have met the criteria for the sampling. The enterprises which have more than 50.000 turnovers are obliged to pay the VAT, and the rest are not obliged to pay the VAT.

- Collection of data

Procedures to collect the data have been organised in that way, where we have consider that is best way to collect data.

- Analysis of data - In the process to estimate data from the survey, an important step is the analysis of information from economical point of view.

During the analysis we found some illogical data as following:

- Expenditure were higher than turnover;
- Wages and salaries for instance in some cases 50 euro/employee
- Turnover per employee 1000 euro whereas wages per employee more than 1200 euros, in such way those figures didn’t make any sense.

In such cases we have used comparative method within the same sector for different enterprises and Comparative method in different time for the same enterprises also.

- Intersection analysis

Example: 500 = purchasing; 400 = Turnover; 300 = salaries; 200 = number of employees

Tab.1

Stat. Units	NACE code	Turnover	Purchasing	Salary	No. of employee	Ratio 4/5	Av. Salary
A	46	3,452,165	2,456,152	75,850	25	0.71	252.83
B	46	16,356,145	10,556,085	95,851	30	0.65	266.25
C	46	15,467,154	10,587,095	85,851	25	0.68	286.17
D	46	12,158,250	17,850,950	72,850	18	1.47	337.27

E	46	8,956,985	6,857,599	10,500	15	0.77	58.33
F	46	589,950	256,355	158,500	27	0.43	489.20
total		56,980,649	48,564,236	499,401	140	0.85	297.26
*taken into account		44,822,399	30,713,286	340,901	98	0.69	289.88

Looking at this we have three cases with illogical data:

Enterprise "D" the cost of buying goods it's higher than Value of Sale or 17,850,950.0 > 12,158,250. In this case we have observed that cost of buying goods has been exaggerated. What we did? We used Average method within the section.

$$\boxed{\overline{X} = \frac{\sum Xi}{\sum Yi} = \frac{448223991}{307132860} = 1.45}$$

Xi = turnover

Yi = purchasing

Based on the result of formula the ratio of Purchasing to Turnover should be 0.69 and not 1.45, after that we corrected the value from 17,850,950.0 to 8,389,192.5 to prove 8,389,192.5/ 12,158,250 = 0.69*100 = 69%

Notice: were from we got the number 8,389,192.5 0.69* 12,158,250 = 8,389,192.5 *euro*

After the adjusted data the table will look like this:

5. Main outcomes of the survey

The SBS provide information about:

- number of employees;
- turnover;
- value of purchases and detail of these purchases;
- value of the inventories at the beginning of the year and at the end of the year;
- value of the taxes paid by enterprises;
- value and details concerning the investment;

This information's is detailed by activities using NACE classification SBS survey as the other STS surveys, get samples from the BR frame. The unique BR identification code is used as key to randomly select units for the sample.

Using the coding system applied in the registration of survey data, each surveyed units is attributed a response code allowing to identify errors both in the surveyed data (in the questionnaire) and in the business frame.

1) respondent unit, questionnaires came back with correct information (full response);

2) total non-response, questionnaires never sent back;

3) data are not useful for estimation;

4) rejected, unknown, moved (blank questionnaires came back);

5) units are ceased, not active, in bankruptcy, etc..

Only errors type 4 and 5 can be associated to a lack of quality in the register. While errors type 5 can be due mainly to the delay in the BR updating process causing over-coverage in the target population that increases sampling errors. Type 4 are errors in the BR that concern identification variables, in particular, localisation variables.

These errors increase both the cost of the survey, the bias (probably the not reached unit is out of the scope for different reasons) and the sampling error (variance) in fact it is a non-response. Both types will be treated as frame errors to measure their impact on sample estimations.

6. Conclusions

The unsatisfactory sampling survey response rate together with the availability of a huge amount of data from administrative sources (balance sheets and tax data) has suggested some adjustments in the SBS production process.

The integration of the original SBS sample with administrative sources has allowed both to increase the response rate and to measure the discrepancies in the final estimation due to unit non-response. Based on that we consider that we have good result for estimation on level of the Country!

Finally we tried to present some techniques which we have used from the beginning of process to the end of this process (sample, collection of data, analysis and the result of the data derived from the SBS survey 2016.

A further analysis on the informative contents of tax data could permit to extend this experiment to other SBS variables. While for other SBS variables which cannot be obtained from administrative sources it will be necessary to develop specific statistical imputation methods. For that aim, it could be desirable that KAS should have an more active role in designing tax forms harmonizing concepts and adding some information useful for statistical purposes. Finally it needs to remark that the use of administrative sources for statistical purpose will imply the continuity and the stability over time of the data flow in order to guarantee the requirements of the Eurostat SBS regulation.

References

1. Bethel, J. (1989), "Sample Allocation in Multivariate Surveys", *Survey Methodology*, 15, pp. 47-57.
2. Deville, J.C. Särndal, C. E.(1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87, pp. 376-382.
3. Garofalo, G. (2000), "Progress Report: The Statistical Information System on Enterprises and Institutions", *14th International Roundtable on Business Survey Frames*.
4. Ohlsson, E. (1995), "Coordination of Samples Using Permanent Random Numbers", in B.G. Cox et al. (eds.*) Business Survey Methods*, New York: Wiley, pp. 153-169.

Thank you for Your attention!

Prishtina; 20.05.2018

Hysni Elshani

Hysni.el@gmail.com
Hysni.elshani@rks-gov.net