Seminar on Registers in Statistics - methodology and quality 21 - 23 May, 2007 Helsinki

Use of the tax data for the purposes of the short-term statistics

Rudi Seljak
Statistical Office of the Republic of Slovenia
rudi.seljak@gov.si

1. Introduction

Through the recent years most of the national statistical institutes have been constantly confronted with the demands for quick and relevant statistical results. All these demands are especially outstanding in the case of the short-term business surveys which are by definition designed to provide quick results with acceptable level of quality. Therefore the short-term statistics is one of those fields of the official statistics where the burden for the reporting and the costs for the producer side represent an essential problem for the statistical institutes. One of the lately most frequently used ways to at least partly reduce these costs and burdens is convenient use of different types of administrative data. Due to the rapid improvement of information infrastructure, these data could be at the disposal timely and in the proper form. These data were originally not collected for the statistical purposes and can many times differ in some methodological aspects but by the use of appropriate procedures they can serve also for the purposes of the statistical production.

In the paper we present the usage of the data, which are provided to Statistical Office of the Republic of Slovenia (SORS) by the TAX authorities and which are originally used for the monthly settlement of the value added tax. In the year 2005 we began to examine the possibilities of usage of these data for the purposes of the estimation of the monthly turnover indices. The wholesale trade activity group was chosen as a kind of "pilot field". In 2005 the feasibility study was carried out and on the basis of the results of this study the fundaments of the new methodology was set up. In the beginning of 2006 we started to regularly produce turnover indices for the wholesale trade, obtained by the new methodology. At the same time we started the feasibility study also for the field of other business services and in the beginning of year 2007 the "new production" of the turnover indices started also for this field.

In the first part of the paper we will present the methodological aspects of the new methodology and describe the statistical process which is of special interest because of the fact that we are merging and processing in the same survey both administrative as well as statistical data. The second part of the paper will be devoted to the description of the special "satellite" register which was set up in order to enable the survey manager the quick and efficient overview over the target population. In this register all the data about the reporting units are stored and updated. The units have special identification and the Statistical Register identification is just one of the atributes in this register. At the end we will shortly describe the main benefits of the new methodology and also point out some problems which should be our chalange for the future work.

2. Methodology

The new methodology for the estimation of the monthly turnover indices uses two types of data. For small number of the largest (according to he turnover) units the data are collected by the classical way, using the post questionnaire which should the units fill and send back to the statistical office till the certain data. These units represent 3% of the whole population in the sense of number of the units but cover more then 50% of the total turnover. For the majority of the units we use the tax authority's data to estimate the monthly turnover. This estimate which is derived out of the items from the tax form is not completely in line with the methodological definition of the turnover and one of the main goals of the feasibility study was to find out if these estimates are good enough to serve our purposes. Therefore in the feasibility study we simulated the new methodology for all the months of years 2003-2005 and then compared the monthly turnover obtained by the new and old methodology. It turned out that the level of the turnover from both sources can sometimes differ essentially but the movement, expressed in the form of the indices, is surprisingly coherent. For the illustration we present the chart where the time series of turnover indices calculated by the new and by the old methodology is presented. The results are from the wholesale trade activity.

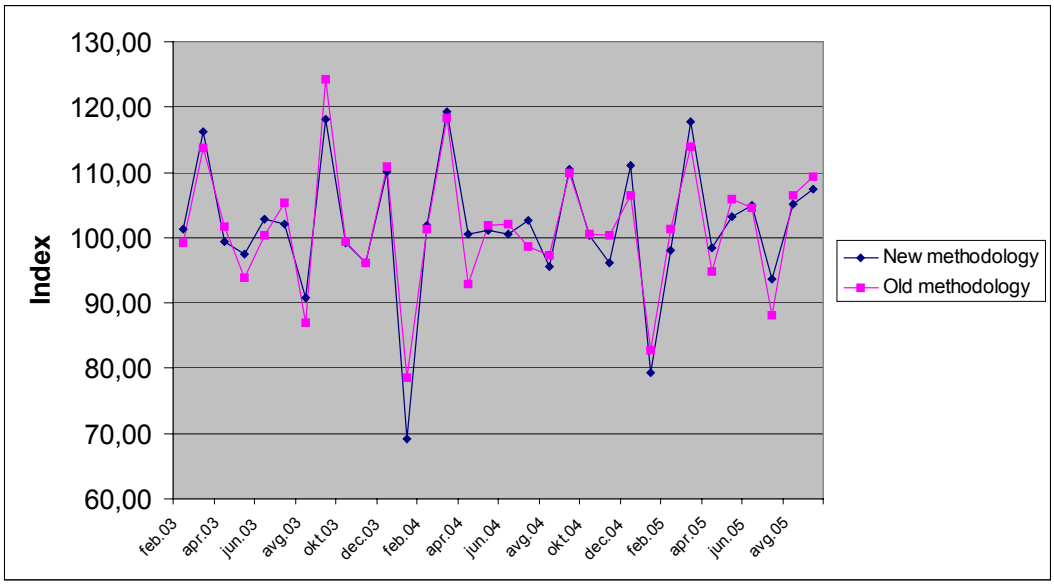

Figure 1: Comparison of time series obtained by two different methodologies

Although generally the tax data can serve well for the purposes of the observation of the turnover movement we detected some cases that could seriously distort the image of the observed phenomena. Such problems usually occur in the case when the enterprise sells the real property. This purchase money is reported to the tax authorities but it shouldn't be included in the turnover. To avoid the serious overestimate of the monthly indices, we had to set up the automatic data editing system which would detect and correct such cases. The system is based on the well known Hidiroglu-Berthelot designed for the cases of the periodical business surveys. With this method the distribution of the month to month turnover change is explored. In the first step the distribution is transform in the way that the transformed distribution is symmetrical. In the second step the extreme values from the tales of the transformed distribution are detected as the outliers. These values are later in the

process re- estimated by the imputation procedure. The procedure should be suitable parameterised and the tuning of the parameters was done during the feasibility study.

3. Statistical process

Since the time gap between the time when we get the tax data and the time when we have to release the results is very short, the system for the processing of the data has to be fully automated and controlled by the survey manager. The whole process could roughly be divided in two parts. In the first part the data from different sources are merged together in the form that enables automated processing. In the second part the statistical processes like editing, imputation and aggregation are executed. Here we will not describe the system in detail, but just present the whole process graphically. The first figure shows the first part of the process.

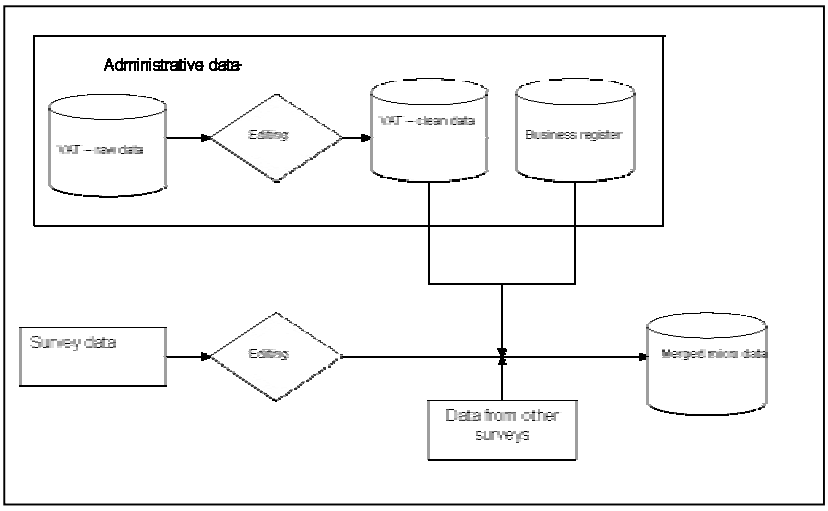

Figure 2: Merging data from different sources

The data which were in the first part loaded in the micro-data data base are then statistically processed:

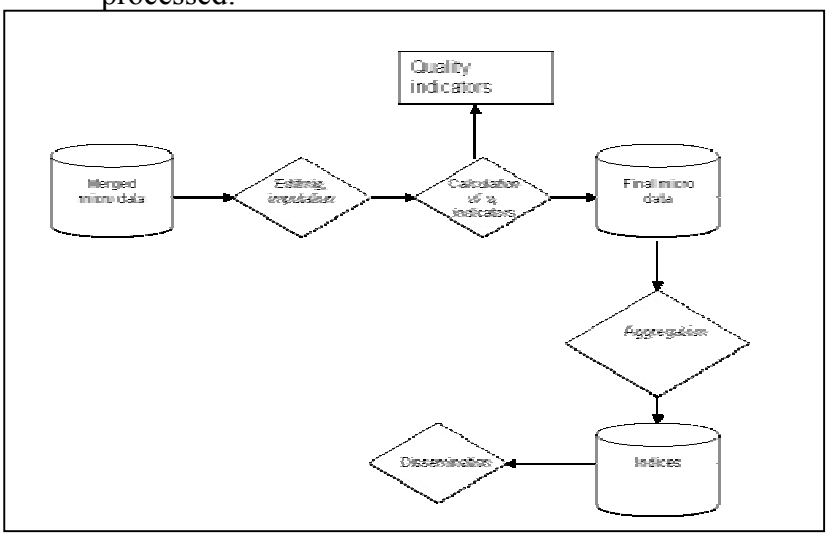

Figure 3: Processing of the data

4. Target population

One of the significant changes in the new methodology was the movement from the random sampling to the cut-off sampling procedure. The reason for this change was most of all the fact that by using the tax data the data for many more units are available and the cut-off procedure would produce much more precise results than the random sampling. On the other hand also the tax data do not cover the whole population of interest. This is due to the fact that the units whose annual turnover is under the certain threshold are not obliged to report their data. Beside this some enterprises that are obliged to report are not obliged to report monthly but quarterly. Due to all these facts it is very important to set up the selection system carefully in order to obtain the target population which would assure the results with the sufficient degree of precision. The main goal of the selection procedure was hence to get the target population which would cover a large part of the population of interest and would, according to the available data, lead to a sufficient response rate. In other words, we would like to avoid too large proportion of the imputed data.

According to the results of the feasibility study, we decided that the target population will be updated twice a year, hence semi-annually. The whole procedure is carried out in two steps. In the firs step the units of the target population are determined and then in the second step the units for which the data will be obtained by the "classical survey", are selected. In the first step the units which fulfil one of the following criteria are selected:

- The semi-annual turnover of the unit is more than 100 000 EUR.
- The semi-annual turnover of the unit is more than 50 000 EUR and the unit has at least 3 employees.
- The unit has at least 6 employees.

For the selection of the "field units" the target population is in each of the activity groups firstly sorted by the descending turnover. Then inside of each of the activity groups so many of the largest units are selected that the share of the turnover of the selected units exceeds the target share of the total turnover. The target share slightly differs between the activity groups but it is generally between 50% and 60%. The number of the selected units which are then surveyed by the post questionnaire is approximately 2% of the whole target population.

As it was already explained it is of great importance for the efficiency of the whole system that we have an effective and transparent system for the management of the set of the observational units. Therefore we decided that we will set up the special satellite register of the units which have ever been included in the set of the observational units. Hence each unit which is included in the target population for the first time, it is also inserted into the register and remains there although it is in some point in time excluded out of the target population. At the time the register consists of approximately 17000 units from wholesale trade and other business services. To enable easier management of the units, especially the management of the demography changes, all the units in the register have the special 6-digit identification number. This identification can remain the same even if the identification from the business register changes. The business register identification is in fact just one of the attributes in this register.

The register plays multiple roles in the whole system. The most important roles are:

- The data on the observational units are stored centrally and the survey manager can access and browse these data easily. Also the data on the units which are not observed at the certain time point are available.
- The survey manager can insert the changes of the data manually or by using the special automated procedure which updates the data in the register with the reported data. By the manual procedures the survey manager can change the data on NACE code, address and the administrative identification. The changes are inserted through the tailor made graphical interfaces.
- The user can see and check all the historical changes. All the changes are namely stored in a special "historical" table.
- By using the graphical interfaces in the access database the survey manager can run and control the individual steps of the statistical process. The procedures are designed as a classical "push the button" system where all the processes are fully automated and the user only needs to insert the set of parameters.

We present the role of the satellite register in the following, a bit simplified picture:

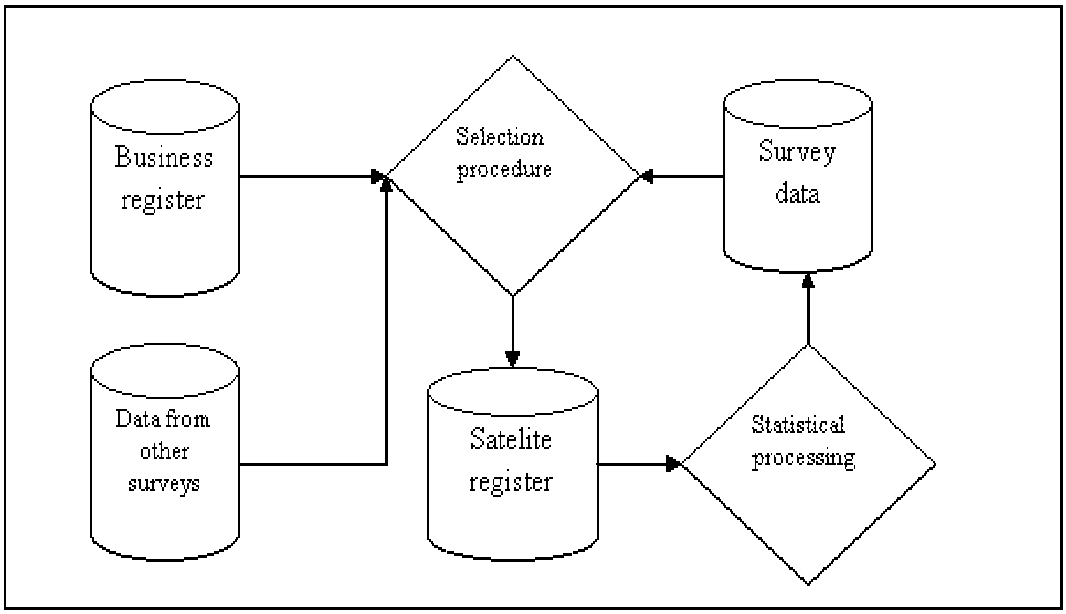

Figure 4: The role of the satellite register in the statistical process

5. Conclusions

SORS started the new methodology for the production of the short-term business statistics in 2006 for the field of the wholesale trade. In year 2007 the new methodology was widened to the area of other business services. The new methodology is based on the use of the tax administration data which the enterprises send to the tax authorities for the purposes of the value-added tax clearance. Although the estimates derived from the tax data are not completely in line with the methodological definition of the turnover, the analyses in the feasibility study shown that they could serve well enough for the purposes of the estimation of the indices. Only small proportion of the largest units in the target population is still surveyed classically by using the post questionnaire. The main reason that these units are still surveyed classically, is that we want at least partially overcome the methodological

differences in the definition of the turnover. The greatest benefit of the new methodology approximately 4000 units were surveyed every month, whereas now only approx. 400 units still get the post questionnaire. There is also considerable cost reduction from the SORS side since the material as well as human resources costs have been significantly reduced.

To enable the survey manager the effective control over the current and historical target population, the access database with multiple functionalities was created. The database could be considered as a special satellite register where the data on the units that have ever been included in the survey are stored and updated. The units in the register have a special identification number while the identification from the business register appears only as an attribute. The database also serves as a central point from where the hole statistical process could be run and controlled. Since the process is fully automated, a lot of outputs are produced with the attention to enable the survey manager at least indirect insight into the "black box" process.

One of the most significant problems of the new methodology are the high leaps in the estimates of the turnover which are usually caused by the sale of the real properties which by the definition shouldn't be included in the turnover. To detect and correct such a cases an automated data editing system based on the Hidiroglu-Berthelot method has been set up. Although the method works sufficiently well there is still a lot of space for the improvements. So the main challenge for the future would be to improve and upgrade the automated data editing system.

One also very important challenge for the future is to introduce the new methodology also for the field of the retail trade activity. The important difference according to other areas is that by the regulation the data from the retail trade are requested much earlier and the tax data at that time is still not available. That's why we plan to produce the first provisional results just with the small cut-off sample of the largest units which will report the data by post questionnaire. Later when we will get the tax data we ill produce and publish all the results again. The plan is to begin the production of the retail trade turnover indices by the new methodology in the beginning of the 2008.

References:

1. Hidiroglou, M.A. and J.M. Berthelot (1986), "Statistical Editing and Imputation for Periodic Business Surveys", Survey Methodology, 12, pp. 73-83
2. L. Lyberg et al. – Survey measurement and Survey Quality, Wiley, 1997
3. Methodology of short-term business statistics, Interpretation and guidelines, European Commission, Luxembourg 2006
4. SORS, Feasibility study on the use of administrative data in the wholesale activity, internal document
5. Council Regulation (EC) No 1165/98 concerning short-term statistics amended by the Regulation (EC) No 1158/2005 of the European Parliament and of the Council