Proceedings of MSP2009
Modernisation of Statistics Production 2009

A way towards optimal use of secondary sources - methodological challenges

Metka Zaletel[1] and Irena Križman[2]

Registers and other administrative sources are nowadays widely used in majority of national statistical institutes in different areas of statistics and for different reasons. The most important reasons are of course the reduction of the response burden, ensuring timely coherent data to the users and decreasing the amount of routine work at national statistical institutes. Besides the influence of using secondary data sources on the statistical process by itself, there is also an influence on the methodological part of the statistical work. The indicator in the European Statistics Code of Practice, "Where European Statistics are based on administrative data, the definitions and concepts used for the administrative purpose must be a good approximation to those required for statistical purposes.", makes national statistical institutes the driving force in the possible harmonization of methodologies used in administration and statistics. In the paper, we show three case studies in the field of population, agriculture and short-term statistics.

Key words: administrative data, registers, methodology, quality

1. Introduction

The pronounced needs of the contemporary users of the statistical results, especially the needs of the political decision-makers, for quick and quality data on different aspects of the modern society reflect in the constantly rising demands for more and more efficient methods of data collection and statistical results production. The results should hence be produced quickly, with low costs and with high quality. Since these demands are most of the time contradictory, the statisticians are faced with the real problem of finding the balance between all these demands of different users.

Together with authorised producers in the national statistical system the Statistical Office of the Republic of Slovenia (hereinafter SORS) in the previous Medium-Term Programme of Statistical Surveys 2003-2007 already adopted as one of its most important long-term objectives the plan of activities that enable as wide as possible use of existing data from various administrative registers and records and the use of innovative methodological and IT approaches. This practice has continued with the next Medium-Term Programme (2008-2012). With continuous activities in this field, producers of national statistics reduce administrative burdens of data providers (enterprises, family farms and persons). SORS and authorised producers of official statistics study the reduction of the burden of reporting units within authorised statistical bodies in Slovenia and abroad and present the results in annual reports on implementing statistical surveys. It is very important to keep the balance between the demands of data users and the requirements for reducing the burden of reporting units. This balance reflects in the awareness that the primary mission of national statistics is to provide quality, comparable, timely and relevant data with the minimum possible costs for everybody involved. Reduction of burdens in statistics can be

[1] Metka Zaletel, deputy Director-General, Statistical Office of the Republic of Slovenia, Vožarski pot 12, SI-1000 Ljubljana, Slovenia, metka.zaletel@gov.si
[2] Irena Križman, Director-General, Statistical Office of the Republic of Slovenia, Vožarski pot 12, SI-1000 Ljubljana, Slovenia, irena.krizman@gov.si

achieved by minimising demands of users (number of variables, frequency of data collection, excluding some of the reporting units) and by using already collected data, either exchanging them within the statistical system or using administrative data with modern technology and organisation.

There are three important statistical areas - population census, agriculture statistics and short-term statistics - where in the past decade administrative sources were shown as the best option for statistical production.

Our point of interest in this paper is methodological correspondence of administrative sources and the quality of these sources. In the second section, a historical development of the usage of administrative sources will be presented. Next, the recommendations set by the European Statistics Code of Practice will be shown. At the end, case studies and lessons learnt will be presented.

2. A bit of history

When discussing the relevance and quality of administrative sources, the history of the development of these sources is very important. As one might expect, the history in various countries is of course different and nowhere as well established as in the Nordic countries. Therefore, we will briefly describe the history of the development of the administrative sources in Slovenia; according to the present situation we could assume that this development could be generalized.

In most of the countries (and again, we do not describe the development of register-based systems in the Nordic countries here), the origins of several registers were developed at national statistical institutes. In Slovenia, the Central Population Register was developed in the 1970s, and the Business Register and the Register of Territorial Units in the 1980s. At that time, NSIs served also as registrars and administrative units, but the majority of NSIs have managed to eliminate the administrative functions. Some NSIs are still serving as registrars, but in general one can conclude that the administrative function of NSIs – at least in the EU – has been excluded or at least separated from the statistical function. The administrative sources with the origin at NSIs in general do not have problems mentioned in the introduction – methodological or quality problems. The reason for that is very simple – these registers were planned to be used for statistical purposes.

On the other hand, many registers and other administrative sources have been developed for purely administrative purposes of the state. During the previous decade, only few of them were used for statistical purposes, but nowadays the majority of them are used in the process of production of statistics mostly because of reducing response burden. Most of them carry with them both important imperfections in the sense of quality and methodological correspondence. The quality and methodological correspondence of these sources depend on the possibility and capability of the national statistical institute to be actively involved in the process of establishing the register (or other data source). In Slovenia, very fruitful results could be shown in various cases when SORS has actively cooperated during the development of the register – e.g. the Real Estate Register, but there are of course some data sources that SORS has not got a chance to be part of the development team.

3. What does the European Statistics Code of Practice say about administrative sources?

It is well known that the European Statistics Code of Practice (Code) was adopted in early 2005; there have been several activities performed to evaluate how well the Code has been adopted across the EU. But our purpose is to examine what are the instructions or directions set by the Code regarding administrative sources. Quick browsing of the Code shows us that the administrative sources appear in three principles:

> (2) Mandate for data collection, more precisely determined by the indicator: "The statistical authority is allowed by national legislation to use administrative records for statistical purposes.",
> (8) Appropriate statistical procedures, more precisely determined by the indicator: "Where European Statistics are based on administrative data, the definitions and concepts used for the administrative purpose must be a good approximation to those required for statistical purposes.", and
> (10) Cost effectiveness, more precisely determined by the indicator: "Proactive efforts are being made to improve the statistical potential of administrative records and avoid costly direct surveys.".

If we summarize, three different issues are tackled with the Code: legal framework, methodological aspects of secondary sources and reduction of burden. In this paper, we are focusing on the second issue – methodological aspects of the secondary data sources; at the same time, our discussion of the main issue will assume that nowadays the usage of administrative sources is a normal way of reducing the response burden and that in the vast majority of countries a proper legal framework enabling the use of administrative sources takes place.

In the case of Slovenia, the legal framework of the Slovenian national statistical system should be pointed out. The National Statistics Act in its Articles 28 and 32 clearly defines the role of SORS as the coordinator of the national statistical system while using administrative data sources for different purposes. Two issues might be raised here as being very important for the efforts to reduce the response burden:

- The possibility to influence the content and structure of administrative sources which enable SORS and other authorised producers to promote standardisation and use of national classification systems on one hand, and on the other hand to introduce necessary variables into the administrative data source and to take care of regular updating of the source where it is feasible to cooperate with other governmental bodies.
- The possibility to use all different kinds of public and private data sources (registers, databases, etc.) free of charge and to link them for statistical and research purposes.

However, as one might expect, the National Statistics Act is not enough to realize the strategies mentioned in the previous chapter. Very pro-active involvement of the representatives of the national statistical system is necessary to establish a stable system at state level which offers data of good quality to the users.

4. Case studies

In this chapter, three different case studies will be shown:

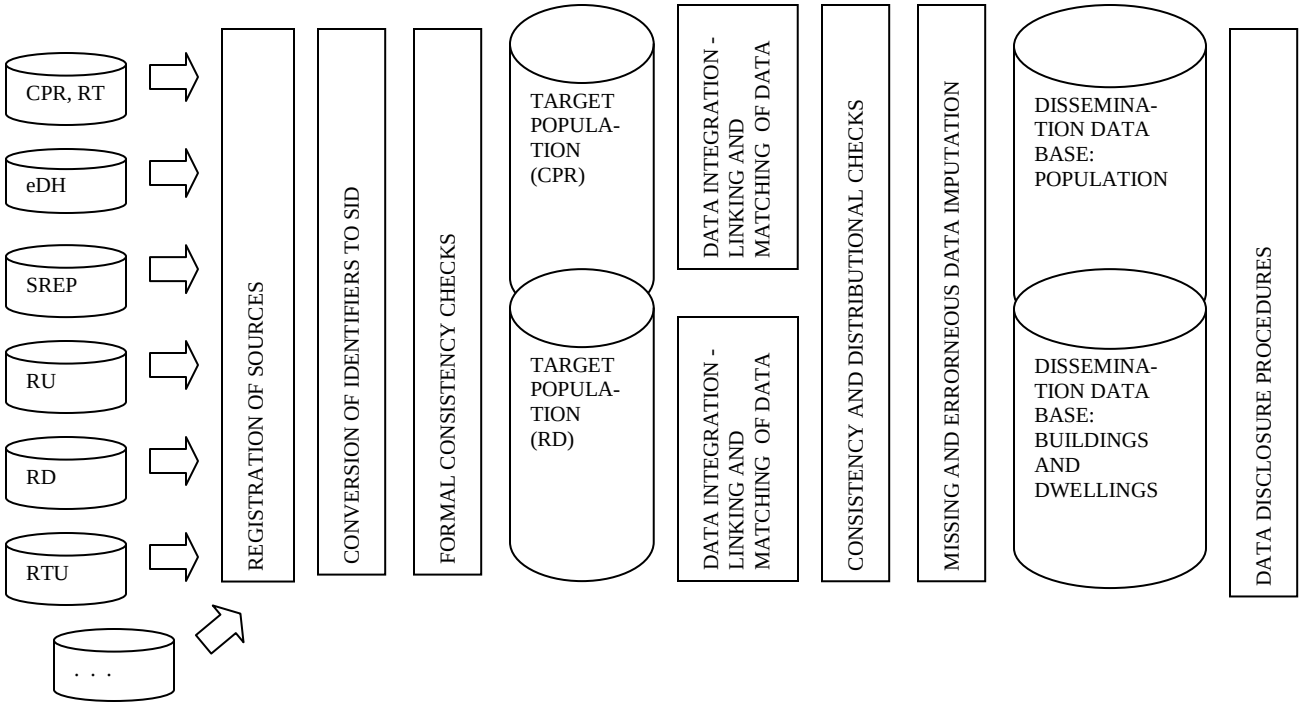
1. Use of administrative sources in the case of a register-based population census where the major source is the Central Population Register. As mentioned before, it used to be developed and for some years also maintained at SORS, therefore it is shown as a case of a high-quality external data source.
2. Use of administrative sources in the case of agriculture statistics where the majority of administrative data sources used have not been developed within SORS, therefore there are a lot of questions concerning the methodology and quality of data.
3. Use of administrative sources in the case of short-term statistics where tax data are used; methodological issues are important in this case.

4.1 Population census

In Slovenia already the 2002 Population Census was partly based on administrative data sources, while the next, 2011 Census is planned to be fully register-based. The quality assessment of the census processes and results is an important obligation of statistical offices.

With a register-based census the 'census data are produced using the method of register estimation, in which several register sources are simultaneously used to define for each statistical unit the value of the relevant variable' . At the Slovenian 2011 register-based census data from three main administrative sources will be used; from the Central Population Register (hereinafter CPR), from the e-Database of Households and from the Register of Dwellings. These sources will be used in order to define the two target populations; the population of persons and the population of dwellings. In addition, other administrative and statistical sources will be used with the aim to cover obligatory and nationally interesting census topics. Other available sources are: the Register of Employment, the Register of Unemployment, the Register of Foreigners, the Register of Territorial Units, the Tax Database, social statistics surveys, etc. In some cases (mostly for the household and family structure) also data from the 2002 Census will be used. Data from other data sources will be linked to the target populations in the data integration phase.

At SORS the following process of a register-based census is foreseen:



CPR – the Central Population Register
RT – the Register of Foreigners
eDH – the eDatabase of Households
SREP – the Statistical Register of Employment
RU – the Register of Unemployment
RD – the Register of Dwellings
RTU – the Register of Territorial Units
... – other sources

In each source relevant variables for the 2011 Census will be identified. For the selected relevant variables the time frame and deviations from the reference date (and hour) will be studied and documented. If there are several sources for the same relevant variable, the priority list of sources will be set.

On the basis of documentation, each source will go through the registration, conversion of the identifiers to the statistical identifier (to a so-called SID) and through the validity checks process. It is necessary that each source is registered, but at that stage the only quality measure which could be implemented is whether it was possible to register the source or not. At conversion of the identifiers to the statistical identifier the number of inadequate records for the conversion could be measured. On that basis the number of records (within the source and in all sources together entering in the register-based census system) which passed through to the following stages of the process could be assessed. With the validity checks the consistency with the predefined range of variables in the source will be checked. Some obvious outliers will also be identified at this stage of the process.

After validity controls and on the basis of rules set (according to the definitions agreed at the international level) the two basic target populations will be set – the population of persons and the population of dwellings. Since the CPR data are, at least if we take into consideration the coverage of persons present on the territory of Slovenia, of a high quality, the definition of usual

residents should not be problematic. The only inconsistencies that could be expected are when the CPR units will be linked to the units from the Register of Foreigners. Some duplicates can occur due to the different purpose of the maintenance of those two registers. At present not much is known on the coverage of the target population in the Register of Dwellings since the mentioned register is in the establishment phase. But the register will be based on the Census of Real Estate which took place in 2006 and the undercoverage could be foreseen only for the newly built buildings and dwellings.

The phase of the data integration could undoubtedly be considered as the crucial stage of the statistical process when quality issues are concerned. While the first part of the quality report, devoted to the relevance of the registers and variables, will more or less be descriptive and textual, the "data integration part" should much more be based on the numerical information. Since this part of the process hasn't yet been covered in the "classical" quality components, there is a clear lack of the standard quality indicators for this phase.

In the integration process several different sources will be merged to a target population list. This merging is planned to be accomplished through the two consecutive steps. In the first step the part of the data source will be merged by using the direct matching by the unique identifier. For the remaining part, which will not be successfully merged through the direct matching process, the statistical matching will be used. Monitoring this process, we can define three matching rates. The overall matching rate provides the rate of all the units (regarding the target population) for which the data from the administrative source have successfully been merged (directly or statistically). The direct matching rate provides the rate of the units merged by direct matching while the statistical matching rate provides the rate of the units merged by statistical matching.

For the variables whose values will be derived from different sources, we can measure consistency of the values from different sources. If the first priority variable source is labelled Y_R and the alternative variable source is labelled Y_A, we can define the consistency rate as the ratio between the number of units for which the condition $|Y_R - Y_A|/Y_R < p$ (e.g. $p = 0.01$) holds and the number of all the units for which we have the data from both sources.

Besides the part of the quality report which will be devoted to the direct monitoring of the statistical process, we will naturally also pay attention to the remaining quality components: comparability and coherence. According to the present plans, the comparability component will mostly be devoted to the comparison of the results of the 2011 register-based census with the results of the 2002 (mostly) conventional census. Special attention will be focused on the investigation of the population structure by the enumeration areas. Since we know that in many cases the administrative residence of the persons differs from the actual (usual) residence, we can eventually face some significant differences in these data.

The main question which should come with the coherence component should be: "Are the results of the register-based survey coherent with the results of some other, "classical" statistical surveys such as the EU-SILC, the LFS and the HBS?" We believe that one of our tasks in the process of the 2011 Census is also to give the user a clear picture about the consequences that the difference in the two approaches (classical census vs. register-based census) could cause.

4.2 Agriculture statistics

In the table below, progress during the last decade of using the administrative data sources in the field of agricultural statistics is shown:

Year	Administrative source	Use at SORS	Activities
before 2001	Lists of agricultural holdings maintained by different institutions	SORS created the sampling frame for each survey separately Pilot Census of Vineyards (data from the Register of Grape and Wine Farms)	
2001	Individual data on agriculture production according to subsidies in 2001	Data directly used in some statistics on agriculture	The form for subsidies was adopted according to the request of SORS (partly adoption of the classification system)
2002	Individual data on agriculture production according to subsidies in 2002	Data directly used in some statistics on agriculture	
2003	Individual data on agriculture production according to subsidies in 2003	Data used to control survey data; also used in the imputation process	Workshop on agricultural statistics and use of administrative data was organized in Slovenia
2004-2005	Individual data on agriculture production according to subsidies in 2004, 2005	Data directly used in some statistics on agriculture	Methodological harmonization between SORS and the Ministry of Agriculture in some variables
2006	Preparation of the Census of Orchards (register based), agreement on cooperation in the field of fishery statistics (SORS – Ministry of Agriculture)	Data on subsidies directly used in some statistics on agriculture Data collection and maintenance of records in fisheries transferred to the Ministry of Agriculture (according to EU legislation)	Transition period (double data collection, quality control)
2007	Direct access to Register of Agricultural Holdings at the Ministry of Agriculture Start of preparation of the Census of Agriculture 2010 (checking of new data sources)	Data used to control survey data; also used in the imputation process Data on subsidies directly used in some statistics on agriculture Register-based Census of Orchards Updating of the Statistical Farm Register directly from the Register of Agricultural Holdings (address part)	

2008	…	At the present time, there are 15 different data sources at the Ministry of Agriculture and its agencies that SORS is able to use.	Continuation of harmonization of classifications, definitions (when necessary)

In this case, there are a lot of methodological and quality issues to solve. For example, there are some variables that are not included in the data on subsidies and hence, questionnaires cannot be abandoned; i.e. purely register-based structural surveys of agricultural holdings are impossible. Besides these, there are also questions with impact on the quality of data:

1. Problems with key-identifier of the agricultural holding cause that approx. 10,000 farms from our register cannot get a proper identifier. This problem is solved manually.
2. Several batteries of variables (e.g. different categories of cattle) can be only partially covered by administrative sources because not all categories are connected to subsidies. Therefore some questions are still needed on the questionnaire (in the case of cattle, four instead of seventeen questions). The question of quality is important here (one part of data is collected in March, another part of the same battery of variables is collected in June); there are several possibilities how to solve it.

4.3 Short-term statistics

Finally, we present the usage of the data, which are provided to SORS by the Tax Authorities and which are originally used for the monthly settlement of the value added tax. In 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 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 2007 the "new production" of the turnover indices started also for this field.

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 a certain threshold are not obliged to report their data. Besides 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 wanted to avoid too large proportion of the imputed data.

The new methodology is mostly based on the usage of the administrative data, but for the smaller part of the units, the data are still obtained by the self administered survey. The main reason for this decision was the intention to overcome the methodological differences in the definition of the

turnover. We also wanted to keep the direct contact with the largest in order to easier control the most important data and the demographic changes in the most important part of the population.

The new methodology for the estimation of the monthly turnover indices uses two types of data. As mentioned before, for the small number of the largest (according to the turnover) units the data are still collected by the "classical way", meaning that the units are surveyed by using the postal questionnaire. These questionnaires should be filled by the units and sent back to the statistical office within a certain deadline. The units that are still surveyed classically represent 3% of the whole population in terms of the number of units, but they cover more than 50% of the total turnover. For the remaining, majority part of the population, we use the tax authority's data to estimate the monthly turnover. These units are hence not contacted by the Statistical Office for the purposes of these surveys. The estimates which are derived out of the items from the tax form are 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 the 2003-2005 period and then compared the monthly turnover obtained by the new and the 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.

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 method 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 transformed in the way that the transformed distribution is symmetrical. In the second step the extreme values from the tails 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 suitably parameterised and the tuning of the parameters was done during the feasibility study.

The use of the administrative data as the main data source demanded considerable changes in the methodological and technical realization of the survey process. The main methodological change is the movement from random sampling to cut-off sampling. Technically the process is completely renewed. Most of the steps in the process are fully automated and could be executed and controlled by the survey manager. The special attention was focused on the editing system for the data originating from the administrative source. Since for these data we are not allowed to check their validity with the reporting unit, we set-up the automated data editing procedure, based on the well-known Hidiroglu-Berthelot method.

The greatest benefit of the new methodology is the response burden reduction: by the old methodology approximately 4,000 units were surveyed, while now only approximately 400 units are requested to report their data every month. There is also a considerable cost reduction from SORS's side since the material as well as human resources costs have been significantly reduced.

5. Lessons learnt and the way forward

In the previous section, we presented the use of administrative sources in three different statistical fields that have to cope with problems of quality, definitions and methodology. From the presented case studies we can draw the following recommendations:

- NSIs should separate the statistical function and the administrative function, but play an active role when establishing new register or other administrative sources in the country.
- Quality issues of registers should be dealt by negotiations with the agency/ministry which maintains the register, which is, of course, the old vision of "editing near source".
- The communication NSI ↔ maintainer of the register/data source should be opened and on a daily basis to result in high quality of data; but nevertheless, NSI should take care of data protection and confidentiality.
- Solutions of methodological challenges/problems are in most cases costly for NSIs, because they usually demand dramatic changes in the statistical process and excellently trained statisticians who are able to solve problems with combination of sources and usage of new statistical methods.
- It is necessary to clarify that reduction of the response burden by using administrative sources in fact generates new demands and needs within national statistical systems.
- Users should get appropriate information on the quality and relevance of the data; it is necessary that all detailed information is available in the quality report.

References

CES Recommendations for the 2010 Censuses of Population and Housing. United Nations, New York and Geneva, 2006.

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

Lyberg L. et al. " Survey measurement and Survey Quality", Wiley, 1997

Methodology of short-term business statistics, Interpretation and guidelines, European Commission, Luxembourg 2006

SORS, Feasibility study on the use of administrative data in the wholesale activity, internal document

Seljak R., Zaletel M., "Tax Data as a Means for the Essential Reduction of the Short-term Surveys Response Burden", Paper presented at the International Conference on Establishment Surveys, Montreal 2007