The impact of the quality of administrative sources on different phases of the statistical process

Zaletel, Metka
National Institute of Public Health
Trubarjeva 2
1000 Ljubljana, Slovenia
Metka.zaletel@ivz-rs.si

Seljak, Rudi
Statistical Office of the Republic of Slovenia
Vožarski pot 12
1000 Ljubljana, Slovenia
Rudi.seljak@gov.si

Introduction

The use of administrative data in the production of official statistics has dramatically increased in the recent years. Although it has to be pointed out that such use has been going on for quite a number of years, it is also a fact that the intention of administrative data use has been essentially changed in the last decades. If in the past administrative data were mostly used for the purposes of sampling frame construction or as the auxiliary variable in the estimation process, it is now more and more popular to use administrative data also as a direct data source. The motivation for such a wide range of activities in this area is most of all the clear possibility of essential budget cuts if the costly data collection in the case of "classical" statistical survey is replaced by much cheaper gathering of the data from administrative records.

As mentioned, the use of administrative sources at different phases of the statistical process has a significant impact on the quality of statistical products and services. First, some general reflections on the use of administrative data for statistical purposes are presented. Then, a cross-word of phases of the statistical process with quality components is shown with a detailed discussion of the impact of administrative sources, threats and opportunities for the national statistical institutes and possible solutions for the weaknesses. Also, the authors discuss the usability of well-known and internationally accepted standards for defining quality components and also the definition of phases of the statistical process.

Administrative data – some general reflections

First of all, it is necessary to point out some of the core issues when discussing the use of administrative data for statistical purposes. The basic questions are relevance and reflection of reality, accuracy of the sources and coherence. These aspects are relevant in all phases of the statistical process.

1. Relevance and reflection of reality

Statisticians should always remember that administrative data were collected for purely administrative purpose and that creators of these data never thought about using them for statistical purposes. Therefore, one should bear in mind that the administrative data present “de iure” world and facts; “de facto” is sometimes far away and surprisingly, sometimes it is quite close. Of course, the main task of statisticians is here to estimate the difference between “de iure” and “de facto” and take it into account.

2. Accuracy

As it is well known, accuracy is the most “historical” dimension of statistical quality since other dimensions were identified as quality dimensions quite late at the end of the previous century. There has been a lot of research on the issue of accuracy, but the vast majority of the research has been devoted to accuracy of data collected for purely statistical purposes. But one should at this point tackle two main points: (1) the accuracy of the administrative data by themselves, and (2) what is the difference between the accuracy of the administrative data and the hypothetical statistical data collected by a questionnaire for the same statistical purpose.

The first question is quite easy to answer for a particular set of data and also guidelines how to do it have already been written in many places. The second question is, of course, very general and could be provoking a wide debate on comparisons of different types of data. Nevertheless, the preference of one or the other data source simply depends on too many too specific factors and it should be studied for each case separately. Here we give only some general considerations which should be taken into account when dealing with this kind of a question. As it was already pointed out, one should be aware that both types of data could be contaminated by errors, but the sources of errors are usually quite different. If in the case of the classical statistical survey the quality of the incoming data largely depends on how good the measurement instruments (questionnaires, appointment letters, etc.) used in the data collection process are and how skilled and experienced the interviewers are, the accuracy of the administrative micro data usually depends on quite different factors. At this point, we would like to point out only two – penalties and benefits for misreporting.

National statistical institutes are using all kinds of administrative data, some of them are more regulated by laws and supported by high penalties in the case of non-reporting (e.g. data from tax authorities), other sources are more “flexible”, their collection is based on some guidelines and directives (e.g. register on bees’ hives) and low or even no penalties are employed. On the other hand, some administrative data are used to determine the benefits for the citizens and therefore these data are exposed to the possibility of wrong reporting with the aim of gaining the benefits. Of course, these cases should be recognized by some procedures at the data collection authority, but it is not always the case.

3. Coherence

The question which is discussed here usually arises when the administrative data are used as the source for the exhaustive surveys where a large number of data items should be collected at the micro level. Namely, in such cases we are usually faced with the situation that the data should be gathered from several different data sources and this can cause all kinds of different integrity and consistency problems. If all these data were gathered with a classical survey, all the needed questions would be included in the questionnaire and it would be much easier to obtain coherence of the data. On the other hand, data from different administrative sources can refer to different reference periods, use different observation units, different target populations and could be based on different conceptual approaches. If such a survey is carried out, it is then of crucial importance that all these differences are studied carefully, that their impact on the quality of final results is minimized as much as possible through the statistical process and that the possible deficiencies derived from such an approach are transparently reported to the users.

Definition of the statistical process

The Joint UNECE/Eurostat/OECD Work Sessions on Statistical Metadata (METIS) prepared the "Generic Statistical Business Process Model" (GSBPM), which was presented at ISI Session 2009 in Durban. The basis for this model was the model used by Statistics New Zealand with slight modifications. The GSBPM is intended to apply to all activities undertaken by producers of official statistics, at both the national and international levels, which result in data outputs. The GSBPM is designed to be independent of the data source, so it can be used for the description and quality assessment of processes based on surveys, censuses, administrative records, and other non-statistical or mixed sources. In short, the administrative data sources, which are our main focus here, are also taken into account.

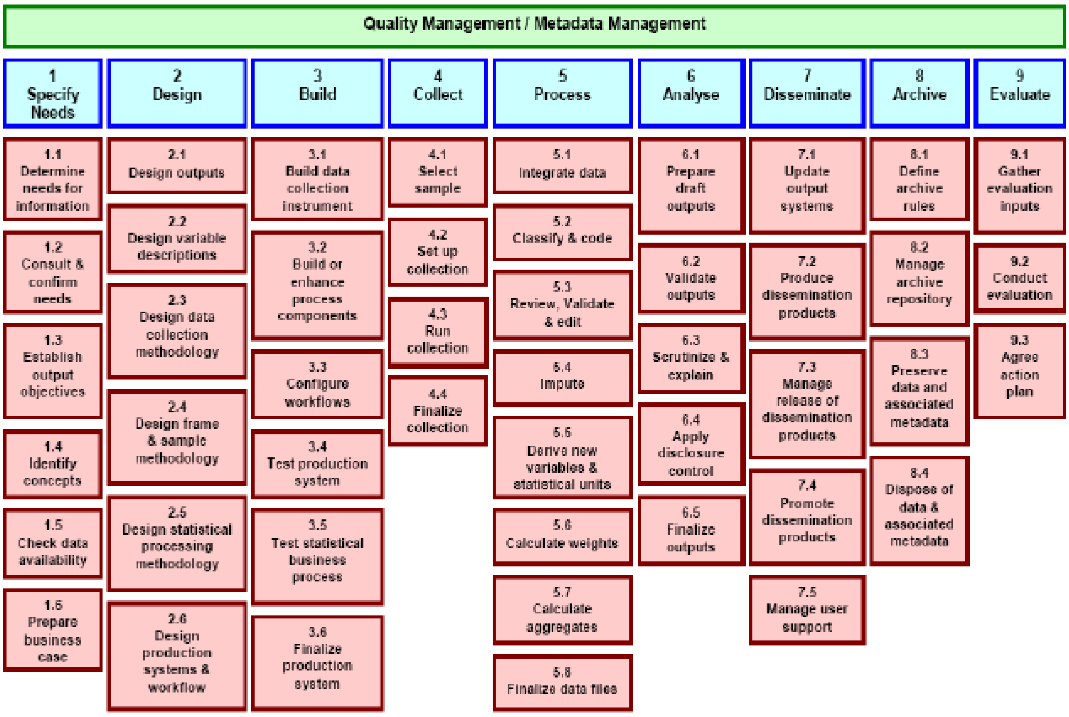
In general, the GSBPM comprises four levels:

- Level 0, the statistical business process
- Level 1, the nine phases of the statistical business process
- Level 2, the sub-processes within each phase
- Level 3, a description of those sub-processes

At level 1, there are nine phases of the statistical business process, which are: specify needs, design, build, collect, process, analyse, disseminate, archive and evaluate. Within each of these nine phases, there are several sub-processes. Each sub-process should have a number of clearly identified attributes, including: input(s), output(s), purpose (value added), owner, guides (manuals and documentation), enablers and feedback loops and mechanism. The figure below shows the whole structure of the business process model.

Nevertheless, one should mention also the two very important overarching processes: quality management and metadata management. Here, some important issues concerning the use of administrative data and the quality of final products could be raised.

Figure 1: A schematic overview of the Generic Statistical Business Process Model


Quality Management / Metadata Management
1 Specify Needs
2 Design
3 Build
4 Collect
5 Process
6 Analyse
7 Disseminate
8 Archive
9 Evaluate
1.1 Determine needs for information
1.2 Consult & confirm needs
1.3 Establish output objectives
1.4 Identify concepts
1.5 Check data availability
1.6 Prepare business case
2.1 Design outputs
2.2 Design variable descriptions
2.3 Design data collection methodology
2.4 Design frame & sample methodology
2.5 Design statistical processing methodology
2.6 Design production systems & workflow
3.1 Build data collection instrument
3.2 Build or enhance process components
3.3 Configure workflows
3.4 Test production system
3.5 Test statistical business process
3.6 Finalize production system
4.1 Select sample
4.2 Set up collection
4.3 Run collection
4.4 Finalize collection
5.1 Integrate data
5.2 Classify & code
5.3 Review, Validate & edit
5.4 Impute
5.5 Derive new variables & statistical units
5.6 Calculate weights
5.7 Calculate aggregates
5.8 Finalize data files
6.1 Prepare draft outputs
6.2 Validate outputs
6.3 Scrutinize & explain
6.4 Apply disclosure control
6.5 Finalize outputs
7.1 Update output systems
7.2 Produce dissemination products
7.3 Manage release of dissemination products
7.4 Promote dissemination products
7.5 Manage user support
8.1 Define archive rules
8.2 Manage archive repository
8.3 Preserve data and associated metadata
8.4 Dispose of data & associated metadata
9.1 Gather evaluation inputs
9.2 Conduct evaluation
9.3 Agree action plan

Importance of different aspects of quality of administrative data in the phases of the statistical process

In the second section, we presented some general remarks regarding the quality issues of administrative data. These general remarks will be upgraded to discussion of the issues of quality according to the different phases of the statistical process. As one can expect, the importance of quality dimensions varies according to the phases of the statistical process. In the table below, we tried to summarize the impact of quality on the process. The reader will notice that "evaluate" and "archive" are missing in the table below; we believe that these two phases are not influenced by the source of data, so we left them out due to the clarity of the table.

	Phases of the statistical process						
Quality dimensions	specify needs	design	build	collect	process	analyse	disseminate
Relevance	X						
Accuracy					X	X	
Timeliness and punctuality		X		X			X
Comparability	X						X
Coherence	X				X		
Accessibility and clarity							

We can conclude that some of the quality dimensions are not influenced by the source of data at all. As an example, we can discuss the "accessibility and clarity". Accessibility and clarity refer to the simplicity and ease with which users can access statistics, with the appropriate supporting information and assistance [1]. Accessibility refers to the physical conditions for users to access the statistical data: where and how it is possible to order data, delivery time, how much it costs (clear pricing policy), access to microdata and metadata, availability in various formats. Clarity refers to the environment in which the data are presented: are data accompanied by appropriate metadata, by graphical presentations, by information on their quality and by information about the extent to which additional assistance is provided by the national statistical institute. We can see that accessibility and clarity are independent of the type of data source and therefore not considered in our analysis.

In the following subsection, we will discuss each phase of the statistical process according to the quality dimensions provided that administrative sources are used.

Phase 1: Specify needs

When planning the statistical survey in the broader sense, considering all possible sources of data – "classical" statistical surveys and administrative data sources – the statistician should take all quality dimensions into account. The most important are: relevance, comparability and coherence.

At this stage, sub-processes of each phase should be considered. We would point out the discussion with stakeholders and checking the data availability as the sub-processes where some issues concern the use of the administrative data. First of all, when the need for information / statistical output arises, the discussion with all possible stakeholders is, of course, necessary and in many countries also obligatory by statistical law. The discussion should also consider the possible sources in the country, availability of administrative sources and their relevance to the problem.

The role of the relevance component in the quality assessment process is significantly changed if the administrative sources are used. In the case of classical surveys it is more or less a product-oriented component, mostly assessing the relevance of the final statistical result in the sense how much it meets the user needs. On the other hand, in the case of the usage of administrative data, relevance becomes a strongly process-oriented component, since a large number of the factors that determine the relevance component derive directly from the first part of the process, when different administrative sources are gathered together in order to be used in the statistical process. In other words, if in the first case the relevance is mostly studied from the perspective of the user, in the second case the relevance component should become a tool for the assessment of appropriateness of the incoming sources for the planned purposes.

There are two aspects of the relevance which should be especially thoroughly studied in the phase when fitness for our purposes is considered: (1) are the methodological concepts that define the variables in the administrative sources sufficiently close to the statistical concepts that are stated in the design of our survey? The fact is that the quality in the case of administrative data usage can predominantly be determined by such conceptual discrepancies. (2) Is the reference period of the variables in the administrative sources compliant with the period targeted by the survey? If there are cases that this is not true, it must be clearly stated in the quality report of the final statistical product.

The next quality dimension relevant to Phase 1–Specify Needs is comparability. Again, when planning the statistical survey in the broader sense, the comparability is crucial. In the case of administrative sources while designing the survey, one can gain from using the administrative source and lose at the same time. On one hand, the administrative source could offer us more data on smaller geographical units, so we can get local (e.g. regional) estimates at a very low cost that could never be achieved by a classical statistical survey. On the other hand, administrative data are always a product of national or even regional legislation; therefore, geographical comparability at the level of countries is questionable. At the end, we have to mention the comparability over time: when designing a survey, all possible shortages of administrative data (e.g. non-statistical definitions of available variables) are taken into account. But the legislation and consequently the administrative data source can in some cases change quite frequently, which causes the changes over the years and also results in non-comparable data. That fact leads us to the first over-arching process, namely quality management. The administrative sources and their legal background should be followed and monitored continuously since a small and at first sight non-visible change can cause dramatically changed statistical results.

The third quality dimension important at Phase 1–Specify Needs is coherence. At the stage of specifying needs, the notion of importance of coherence is important; statisticians can at this point of the process enable the results which are coherent. If we are considering coherence only in the sense of coherence with statistical results from other areas (e.g. national accounts), the influence of administrative data usage can be both positive as well as negative. In the case that different statistical surveys "operating" the same (or at least similar) area are using the same administrative source, this should increase the level of coherence of the results. On the other hand, if in one survey the administrative and in the other statistical source are used, the impact on the coherence could be exactly the opposite. But even in the latter case, this shortcoming could be turned into an advantage if the data from two surveys are properly combined in order to increase the quality. For example, if in a structural survey an exhaustive field survey is employed, these data could be used to overcome some eventual imperfections in the short-term survey which is based on the administrative data.

Phase 2: Design

When discussing the importance of quality components of administrative sources in Phase 2–Design of the survey, we can realize that the most important dimension is timeliness. Timeliness reflects the length of time between the period when the statistical phenomenon was observed and the release date of data. At this stage, we would like to point out the importance of distinguishing between registers and other administrative data sources. Registers are usually up-to-date and their transfer to the national statistical institute can be done quite soon after the reference period. On the other hand, other administrative data sources are collected on an annual basis, quite often the data collection takes place quite late after the reference period. Then, some administrative procedures usually take place (cleaning the data, preparation of some decision, period to accept complaints from citizens, etc.) and therefore, the national statistical institute receives the data quite late. This is the main reason why this quality dimension is very important at the phase of design of the survey: statisticians have to take different time lags into account to achieve the requested timeliness. Unfortunately, from the experience most of the European statisticians can conclude that using of administrative data takes much longer, but it is cheaper.

For several years SORS has had the right to get from the tax authorities all the personal income tax data. Hence there is no longer any need to collect these data in the "classical surveys" and there are several significant advantages arising from this fact, especially the following two:

- Data from the tax office are believed to be much more accurate than the survey data since it is well known that the income data are sensible data and very frequently subject to reporting errors.
- Data are on disposal for the whole population, which enables their usage for the calibration purposes, which can also increase the accuracy of the results.

On the other hand, there is a problem with the timeliness of these data. The tax office needs some time to collect the data for the whole population, process and verify them, and then deliver them to our office, where they could be further processed. The final tax data are usually provided more than one year after the end of the reference year. For the EU-SILC survey, which is the main user of these data, the estimated delay of the release, due to the delay of the tax data, is approximately 10 months.

Phase 4: Collect

Also in Phase 4–Data collection, the timeliness is the most important quality dimension. Phase 4 is quite short when using exclusively administrative sources, but there are other considerations concerning timeliness described above.

In the case of the Agriculture Census, carried out by SORS in 2010, only part of the data was gathered with the classical survey, by using a CAPI questionnaire, while the other (quite significant) part of the data was gathered from several administrative sources. Use of all the available and relevant already existing sources significantly eased the response burden and reduced the survey costs. The questionnaire was much shorter than it would be without the complementary data sources, which certainly improved the quality of the collected "field data". On the other hand, usage of the different data sources increased the need for more expert work on data integration and data editing, carried out inside the office. A lot of effort has been put in the editing phase to overcome all the inconsistencies caused by the usage of several sources. Also some administrative data were provided late and some results could probably be disseminated earlier if all the data were gathered in the field. However, it is clear that all the benefits of the administrative data usage by far outweigh the above mentioned deficiencies.

Phase 5: Process and Phase 6: Analyse

The "Process" and "Analyse" phases can be iterative and parallel. Analysis can reveal a broader understanding of the data, which might make it apparent that additional processing is needed. Activities within the "Process" and "Analyse" phases may commence before the "Collect" phase is completed. This enables the compilation of provisional results where timeliness is an important concern for users, and increases the time available for analysis. The key difference between these two phases is that "Process" concerns transformations of microdata, whereas "Analyse" concerns the further treatment of statistical aggregates (see [3]). These phases are the most demanding when a combination of statistical and administrative sources is involved. Sub-processes such as integrate data, classify & code, review, validate & edit, calculate weights, etc., are demanding and complicated. In these phases, accuracy and coherence are exposed as the important quality dimensions. As one could notice, accuracy has not been mentioned in the earlier phases of the statistical process, but at this stage it becomes of the ultimate importance. In the case of classical surveys the accuracy component, usually presented through different types of errors, is by all means theoretically the most precisely defined and described component. Terms such as sampling error, non-response error, measurement error, etc., are well known to all the persons handling the execution of the statistical surveys. However, there is a clear lack of the strong and consistent framework for the cases where the administrative or combined data sources are used. Here we give just a few reflections about the adaptations that should be taken into account when the well known error-categories are considered in such cases.

(1) In most of the cases when the decision is made to use the administrative data, these data are available for the large part of the target population and therefore the sampling approach is not sensible in such cases. Hence, sampling errors are rarely present in the case of administrative data usage. However, many times the price of getting rid of sampling errors is the increase in the bias in our results. The main source of the bias usually comes from the coverage problem of the administrative source or sometimes from the problems with the complaints of the reference dates. Although it is not an easy task, an effort should be made to at least approximately estimate the bias derived directly from the fact that the administrative source is used.

(2) The measurement error is an especially "difficult to assess" component in the case of administrative data. Since in such cases the collection process is separated from the statistical process, we are usually limited by the editing procedures aimed at finding the erroneous or suspicious data, but the verification of these data at the data source is usually not possible. In these cases the close cooperation with the data provider is of crucial importance. Namely, the data

provider (administrative authority) can (besides the data themselves) sometimes provide useful information which can then be used for the quality assessment purposes. In any case we should avoid the temptation of considering the data coming from the administrative source as being of such high quality that no additional data editing is needed.

(3) The concept of non-response can be quite ambiguous in the cases when just the administrative data are used in the survey. In such cases it is usually difficult to distinguish it from the concept of coverage error. Let us assume the "classical" situation when we have a list of units determined in advance which represents our target population and then one or more administrative sources are merged (using direct or indirect linkage approach) to this list. It is mostly inevitable that after the integration process there are units for which some or even all target variables are missing. And it is in most cases difficult to know whether they are they missing due to the imperfections in the data collection phase or due to the different target population of the statistical and administrative purposes.

In Phase 1–Specify Needs we already discussed the coherence and raised several issues. Here, the same issues are on the table again, but in Phase 1 the decision about the design was made and now the implementation of the design is crucial.

In the calculation of the monthly turnover indices for the services sector, a combination of the survey data collected by the questionnaire and the VAT data provided by the tax authorities is used. The field data are collected only from a small part of the large units, while for the smaller units VAT data are used. The large units represent 3% of the whole population in the sense of the number of units, but they cover more than 50% of the total turnover. The problem is that from the turnover calculated out of the tax data tax is not always completely in line with the methodological definition of the turnover. Therefore, a special editing process was set-up to correct at least most significant of these discrepancies. Hence in this case the large benefits in terms of the response burden and cost reduction are sometimes offset by the smaller accuracy of the data for small units.

Phase 7: Disseminate

When discussing the importance of quality components of administrative sources in Phase 7–Disseminate, we believe that comparability and timeliness of administrative data have the major influence on this phase. Both issues were described at the earlier phase of the statistical process, but at this stage there is an opportunity to explain and show the advantages and disadvantages of administrative data usage. As mentioned, in most cases the use of administrative sources means longer periods to publish the data and fewer possibilities for international comparisons. Phase 7 is the ultimate time to present the important methodological challenges to advanced users, to explain the key issues to researchers using microdata and to properly disseminate the results to the general public. The fact is that the general public is not very interested whether the results are compiled from the data gathered in the classical survey or from the administrative source, but more or less expects accurate, comparable data in short time. On the other hand, the advanced users do need to know which data source was used, what the advantages and disadvantages of the used data source are and what the consequences of the selected source for the quality of final results (or microdata) are.

Conclusions

In the paper we tried to highlight the key issues in the crossword of the phases of the statistical process and the quality dimensions while using the administrative data source as an exclusive or supplementary data source for the production of the statistical results. While the generalised statistical business process model is designed for the surveys independent of the data source, the quality assessment framework is mostly tailored for the "classical" surveys. In the paper, we wanted to point out the importance of some quality issues which arises in different phases of the statistical process. In reality, many national statistical institutes are aware of the important issues such as different definitions of the variables, but on the other hand, not a lot of efforts are put into phases of specifying needs and disseminating the results. One could expect that these two phases are similar as in classical surveys, but some attention should be focused on these activities, too.

REFERENCES

[1] ESS Standard for Quality Reports, Eurostat, 2009.

[2] Wallgren A., Wallgren B.: Register-based Statistics; Administrative Data for Statistical Purposes: John Wiley & sons, 2007. Available at:
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf

[3] UNECE Secretariat: Generic Statistical Business Process Model, Version 4.0 – April 2009. Available at http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

[4] 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

[5] Lyberg L. et al.: Survey Measurement and Survey Quality, Wiley, 1997.