Checking the Usefulness and Initial Quality of Administrative Data

Verschaeren, Frank
Statistics Belgium, Databases and Nomenclatures
Simon Bolivarlaan 30
1000 Brussels, Belgium
E-mail: frank.verschaeren@economie.fgov.be

1. Introduction

Administrative data is increasingly being used in the production of statistics as an alternative to or a replacement of survey data. In Finland for example, under its statistical act the compilation of statistics must always in the first instance make use of data collected in another context. Direct data collection is only resorted to if the information is not otherwise available. Statistics Finland obtains most – some 95% - of the information it needs for producing statistics from existing administrative data (Leivo, 2010).

From a historical point of view, relying on administrative data is nothing new at all: the first formal offices for official statistics were frequently named “Table offices”, reflecting the fact that their purpose was to summarize administrative micro data into *tables* of macro data (Nordbotten, 2010).

Reducing the burden for respondents is named as the main driver behind this renewed interest in administrative data (Daas et al., 2007), and reuse of existing pieces of information for different purposes effectively permits this burden reduction. At the same time, with the advancements in data storage capacity, processing speed and communication facilities during the last decades, the high volume of administrative records can also play an important role for statistics in contributing to the realization of a clearinghouse architecture (Sundgren, 2004) where micro objects of the society, such as an individual, enterprise or institution are considered as interrelated and interacting elements of a system. Micro data from different sources are brought together in a data space where they can be combined to produce new statistical products.

Administrative data have an advantage over survey data in that they cover a large part or sometimes even the whole of the population, allowing thus to fill in data values for a large number of elements in the “data space” or cube. However, using these administrative data compared with use of survey data requires more attention since the administrative data holder has not usually tailored his collection in accordance with statistical concepts, standards and requirements. A thorough evaluation of the usability of the data will have to precede the incorporation of the data in the data space. Once an administrative data source is withheld for further use in statistical production, the data may need statistical preprocessing to solve conceptual and matching problems before they can be used. It might be necessary to transform the incoming data in order to make them compatible with those already in accordance with statistical concepts.

Statistical offices monitor their incoming survey data and have a collection of procedures in place to guarantee and improve the quality of these data. Examples of these are the pre-testing of questionnaires, training of interviewers or other persons involved in data collection, reviewing response data for unexpected results and unusual patterns, and conducting evaluation studies. Like any secondary source, administrative data are not collected by the statistical institute; the collection process can be very different. It is in general in the hands of other public administrations that have a specific objective to fulfill, like for example taxation. As a consequence, some types of information in the administrative dataset might have received very much attention while other variables are out of the main focus and of a more uneven quality. In some cases reported values will never be changed by the administrative data holder, even if they are known to be wrong.

As is the case with survey data, individual administrative data need to be looked at before clearance can be given for further use in statistical production. The same type of errors that appear in survey data can be found in administrative data sets (Bakker et al., 2008), depending on how those secondary data have been collected, checked and updated by the administration. Hence the importance of having procedures in place to detect and resolve data quality issues with incoming administrative data.

Supporting members of the European Statistical System (ESS) in the implementation of a more efficient way of producing enterprise and trade statistics is one of the priorities set forth via the Modernisation of European Enterprise and Trade Statistics programme (MEETS). Moving from direct data collection to re-use of existing administrative data is considered to play a major role: the information already collected for administrative purposes by other institutions should be used to the greatest possible extend, with the aim of reducing the reporting burden on businesses and to improve the quality of statistical information.

The use of administrative data for business statistics has country specific problems as well as problems common for most of the statistical institutes. The common problems concern the methods of quality checking, editing and estimation of missing variables. One of the ways to help National Statistical Institutes (NSI's) in Europe solve such common problems is to create a European Statistical System Network (ESSnet) where several Member States interested in the topic can collaborate on the common task, and then disseminate the results to non-participating members. An ESSnet on the use of administrative and accounts data in business statistics was set up to address these problems.

One of the work packages (WP2) of this ESSnet aims to:

- Help statistical institutes in considering all relevant issues in checking the usefulness of new (or changed) administrative data sources.
- Examine the practices of the Member States on the initial checking of administrative data and produce a list of recommendations.

2. Current situation

WP2 started in the second half of 2010, more than a year later than most of the other work packages in the ESSnet. By that time a first stocktaking exercise had been held on existing practices in the uses of administrative data for business statistics (WP1) and existing methods used for quality assessment for statistics that are fully or partly based on administrative data. (WP6, Quality indicators)

The information that was gathered in this was, could be used to make a first assessment of the state of affairs and to identify areas where exchange of good practices would be most welcomed.

In the first part of the WP's work a checklist was prepared that helps *considering relevant issues before acquiring a new administrative data source or if an existing data source is to be changed*. We predominantly relied on the work that has been done by Statistics Netherlands (SN): most of the quality studies performed by other NSI's have focused either on the quality of data collected by surveys or on the quality of the statistics produced. Metadata related quality aspects of secondary data sources received less attention although data are useless without a good understanding of their accompanying metadata (Daas and Ossen, 2011). The Dutch experience shows that the use of a checklist for evaluating the potential usefulness of secondary data sources has a number of advantages:

a) It provides a structured way of looking, assuring that the user has paid sufficient attention to the preconditions that are known to be of great importance.

b) Gathering the essential information needed to complete the checklist demands a limited effort compared to evaluating the quality of the actual data. Needless and time consuming efforts can be avoided if

factors that block the use of the data are detected from the start.

c) The results can show where further clarifications are needed, or which topics should be discussed with the data holder.

The draft version of the checklist guides the user through a limited number of general questions like the name of the data source and the administrative data holder's contact information. After that, questions are asked about the data content. A short description of the most important units, variables and events should be given, together with information on unique keys and time references.

When the outcome of the content part of the checklist is negative, the evaluation can be halted: there is no need to go further. Otherwise, a third and last section asks about delivery related information. A comparison is made between the needs of the NSI and the options that are available from the data holder. Costs and legal aspects are also part of the delivery aspect.

The information about "not really useful" data should not be discarded: the intended use of a data source is one of the parameters in the evaluation. Other interested potential users of the data could be looking from a different angle and arrive at different conclusions. Having the possibility to browse existing checklists will at least provide insight in what is basically known about the data sources that were already examined.

Administrative data providers produce data primarily for their own use, they seldom benefit from sharing the data. NSI's have little to offer in return, and are usually in a weak position to discuss the method of transferring the data: most of the time they are glad just to receive the data in any format. Nevertheless, receiving the data in the right way (how, when and in what condition) can be crucial for the NSI's statistical production.

NSI's have to take into account the dependency from external partners and their vulnerability for changes in the databases from where they obtain secondary data. Creating and keeping a good working relation with the administrative data holder is recommended as a good practice: even in cases where the NSI has little influence, it offers possibilities to discuss planned changes and to inform the administration about consequences of those changes. The ESSnet's WP2 will collect and analyze country experiences, and try to formulate further recommendations.

Vulnerability for changes is to some extend also related to how transmission and storage of the data are organized (Koskinen, 2010). Systems built as ad hoc solutions can work very well for a particular dataset, but may turn out to be inflexible when the source changes. Heterogeneity in the transmission and in the handling of incoming datasets can become problematic when more and more administrative data -often in large volumes- are used. But most important is that issues caused by changes in raw data can be isolated. This is the first step in resolving the problems caused by these issues.

Even if standardization of receipt and handling of incoming administrative data cannot always be achieved, NSI's should have a clear view of what they would like to obtain before starting negotiations with a data holder. WP2 will offer support by highlighting the aspects to take into consideration, illustrated by country examples.

Standardizing transmission and storage of administrative data opens up the way for the common checking of incoming data at a unique entry point as part of a more general data quality management. This would fit in the vision of a modernized statistical production system, based on a holistic approach rather than a fragmented one (Radermacher, 2010). Duplication of work by different users in the statistical office would be minimized, and coherence of the data would be better guaranteed in this approach.

The stocktaking exercise on the use of administrative data for business statistics and on existing methods used for quality assessment for statistics that took place before the start of this WP, showed that administrative data are widely used; that quality issues are considered important, and that NSI's check their administrative data. But: even if information is checked and even if these checks are quantitative in nature, the checks are normally made as part of the research process and are not necessarily carried out on a regular or formal basis: there is in many cases no overall approach to managing administrative data quality issues in the NSI.

WP2's objective is to present a common approach and good practices that should help to reduce the resource needed for the production of business statistics. A reference document will be produced with techniques and guidelines for ensuring data quality, generic enough to be applicable in the different NSI's. It is expected that the sharing of good practices will contribute to a more efficient, transparent and harmonized use of administrative data in official statistics.

Finding and resolving data quality issues do in many cases require domain expert knowledge while administrative datasets can be very heterogeneous in form and content within and between countries. Similarities between sources in different countries can be found where European legislative acts come in to play, a good example is Value Added Tax (VAT). In order to find good practices that are applicable in a country-specific context, the work will mirror this duality. General quality control procedures will be presented from a data-centric view on the topic, making abstraction from what is domain specific for the data source. A separate work stream starts from the opposite side, looking at three domains that have enough resemblances in the different ESS countries and are considered very important as a source of administrative data: VAT, employment and business accounts.

The data-centric approach starts from analyzing both data and metadata. Data profiling (also referred to as data discovery) will generate accurate metadata as an output of the process by relying on the data for reverse-engineering the metadata and comparing it to the metadata offered by the administrative data holder or kept in the NSI. The real metadata can then be used to calculate the violations of the metadata in the dataset. Visualization techniques could be used to detect anomalies that otherwise would pass the testing.

Efficiency gains could be made by automating the process and repeating the assertive testing (checking against the data rules) for a set of relevant checks, thus creating the possibility to track changes in data quality over a period of time.

Discovering anomalies while checking the data is only the first step, investigating the causes, developing and implementing remedies and monitoring the results are all essential components of good data quality management. This is more than just cleaning data; it is also about preventing errors to occur in the future. Providing feedback tot the administrative data holder on type and number of specific data errors for example could lead to adaptation of forms, procedures or software at the level of the administration, improving the overall quality of the data received by the NSI.

To illustrate the domain based approach, a short overview is given of the work that has been done for detecting errors in VAT turnover.

Turnover from VAT is one of the main administrative sources used in the production of official statistics, so it is no surprise that several methods for detecting errors could be identified in a literature review on international best practice in this area. The most promising methods were described and the effectiveness of each of these methods was evaluated, at this time with VAT data from one NSI. We have used a number of diagnostics to allow for some comparison between methods, especially in relation to the methods for detecting suspiciously large and suspiciously small Turnover values. These include the proportion of businesses identified as suspicious within each industry and size class and the average size (employment) and VAT Turnover of suspicious businesses compared with the rest of the class. If an independent measure of Turnover is available, it should be possible to estimate the effectiveness of the methods in detecting errors.

However, care should be taken over differences in definitions of variables. For example, survey Turnover is often defined differently to VAT Turnover. In this case it would not be valid to conclude that a business with VAT Turnover different to survey Turnover in the same period is an error. However, if a business is detected as an error and has VAT Turnover similar to the survey Turnover, then this can be viewed as being a likely 'false hit'.

Suspicious patterns that are commonly observed in VAT Turnover data are also described. These are easily identified and generally imply that the business has made a reporting error. It is recommended that VAT data are checked for these patterns before implementing any other error detection method.

Another type of error that is relatively easy to identify and correct is the unit error. If businesses are used to reporting Turnover in thousands of Euros, they may continue to report in this way even if VAT Turnover should be reported in Euros. Left untreated, unit errors are often the largest errors in Turnover data. It is recommended that an automatic method is developed to detect and correct any unit errors in VAT Turnover data, before applying any other rules to detect suspicious Turnover values.

A list of seven key principles for editing administrative data, largely inspired on work done by Statistics New Zealand (Seyb, 2009) is put forward:

1. Maintain the original data supplied by the tax office whenever possible.
2. Choose methods that take account of all statistical uses of the data – macro and micro level if appropriate.
3. Make good use of historical and auxiliary data.
4. Automate the process of detecting and correcting errors in VAT Turnover data as much as possible, to ensure best use of resources and consistent results.
5. Keep an audit trail for any changes to VAT Turnover data – this should include keeping the original (unaltered) data and producing diagnostics for the detection and correction process.
6. The implementation of automated detection and correction methods should be flexible for future improvements.
7. Make every effort to understand how the VAT Turnover data are returned and processed as this is a useful indicator of potential sources of error. Good dialogue with the tax office is essential for this.

Preliminary conclusions

If the current situation was to be described in one word, that word would be "fragmentation". Although the use of administrative data is widespread, and quality is considered important, much of the data quality management efforts are either made on an ad hoc basis, or aimed at one specific aspect of data quality.

WP2 shows that there is potential for a more comprehensive approach, and has produced a tentative outline that allows for more integration of different aspects (like detecting, resolving, and preventing issues). The outline is however only a working document, it has to be challenged, validated and transformed into a practical guide for users in statistical offices.

Checks, recommendations and good practices will be further developed in close cooperation with several members of the ESS.

REFERENCES

Bakker,B.F.M., Linder, F., van Roon, D. (2008). Could that be true? Methodological issues when deriving educational attainment from different administrative datasources and surveys. IAOS Conference on Reshaping Official Statistics. Shanghai, October 2008.

Daas, P.J.H., Fonville, T.C. (2007). Quality control of Dutch administrative registers: An inventory of quality aspects. Seminar on Registers in Statistics – methodology and quality. Helsinki, 2007.

Daas P.J.H., Arends-Toth J., Schouten B.,Kuijvenhoven L. (2008). Quality framework for the evaluation of administrative data, Q2008 conference on Quality in Official Statistics, Rome, July 2008.

Daas, P.J.H., Ossen, S.J.L. (2011). Metadata Quality Evaluation of Secondary Data Sources. International Journal for Quality Research, accepted for publication.

Koskinen, V. (2010). Managing administrative data in statistics, Q2010 conference on Quality in Official Statistics, Helsinki, May 2010.

Leivo J. (2010). The Developing Business Data Collection and measuring Response Burden, Q2010 conference on Quality in Official Statistics, Helsinki, May 2010.

Nordbotten, S. (2010). The use of administrative data in official statistics - past, present and future : with special reference to the Nordic countries. In: Michael Carlson, Hans Nyquist, Mattias Villani (red.), Official statistics : methodology and applications in honour of Daniel Thorburn, pp. 205-223, Stockholm.

Radermacher, W., Barcellan, R. (2010) The European Statistical System's reaction to the statistical consequences of the financial crisis, IFC Bulletin No 33, 08/2010, p 371.

Seyb, A., Stewart, J., Chiang, G., Tinkler, I., Kupferman, L., Cox, V. and Allan, D. (2009). "Automated editing and imputation system for administrative financial data in New Zealand", Proceedings of UNECE Work Session on Statistical Data Editing, Neuchatel, October 2009.

Sundgren B. (2004). Statistical systems: some fundamentals, Statistics Sweden, September 2004.

ABSTRACT

The use of administrative data in the production of business statistics has become more and more widespread. Statistical offices can obtain efficiency gains by the re-use of data from external sources, and counter the increase of non response in sample surveys with this strategy.

However, efforts are needed to ensure the quality of the data, because very often the administrative and other external data are not available in the form needed for statistics.

Since working with administrative data sets is still relatively new and the technology behind it changes over time, a systematic approach to qualifying the usefulness of administrative sources is only just emerging. Finding and resolving data quality issues with incoming data is in many cases performed on an ad hoc basis rather than by a repeatable set of processes.

One of the work packages of the European Statistical System Network (ESSnet) project on the use of administrative data in business statistics aims to assist statistical offices by collecting and presenting experiences with selecting and preparing administrative data for use in the production of statistical information.

The paper will present the first results achieved in elaborating a checklist for evaluating the usefulness of administrative data sources and on finding and resolving initial data quality issues in data sets before they are used further up in the chain..