Checking the Usefulness and Initial Quality of Administrative Data

Frank Verschaeren

on behalf of Work Package 2 of ESSnet on Admin Data,
email: frank.verschaeren@economie.fgov.be

Abstract

Work Package 2 (WP2) of the ESSnet on the use of administrative and accounts data for business statistics aims to provide guidelines for NSI's examining the quality of administrative data as input for the statistical production process.
The work package is set up to meet two objectives:

- To help Member States examining the usefulness of available administrative data for business statistics
- To help Member States for checking initial quality of administrative data before introducing it into statistical data base

A checklist is being developed that helps NSI's consider all the relevant issues in evaluating the usefulness of administrative data before a new source is acquired or if an existing one is to be changed.
Once the data become available, efforts are needed to ensure the quality of the data, because very often the administrative data are not fully compliant with statistical needs. This WP looks closer into methods of detecting and resolving quality issues at the initial stage of receiving the data.
The paper will give an overview of the work planned until June 2013 and present the first results of this WP in elaborating a checklist and on methods for investigating initial data quality issues.

1. Introduction

The European Statistical System network (ESSnet) on the uses of administrative and accounts data for business statistics began in 2009 and is part of Eurostat's Modernisation of European Enterprise and Trade Statistics (MEETS) programme. The ESSnet is a collaboration between the National Statistical Institutes (NSIs) of European Member States to develop best practice recommendations for the uses of administrative data. The ESSnet consists of nine work packages (WP), which address the following:

WP0 – Management and administration
WP1 – Uses of administrative data
WP2 – Checklist for administrative data
WP3 – Methods of estimation for variables
WP4 – Timeliness of administrative data
WP6 – Development of quality indicators
WP7 – Statistics and accounting standards
WP8 – Creation of an Information Centre (http://essnet.admindata.eu/)
WP9 – Training and exchange of best practice

Work Package 2a (WP2a) of the ESSnet on Admin Data focuses on the quality in the sense of 'fitness for use' of administrative data as an input for the production of statistical information. The first question to answer is whether it is OK to start using the data at all.
In an ideal situation, NSI's would be able to fall back on a standard instrument or procedure to evaluate new data sources. WP2a builds further on the work that was already done by Statistics Netherlands[1], and developed a checklist that could help statistical institutes in considering all relevant issues when examining the usefulness of new (or changed) administrative data sources.
Administrations have their specific objective to fulfil, 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. With the large volume of data arriving at the NSI, it is imperative to inspect the data in an organised way, and to resolve data issues as quickly as possible. Work Package 2b

(WP2b) was set up to formulate guidelines and recommendations for NSI's checking these data.
WP2 began in the second half of 2010, a year later than the other Work Packages. In this way, it could profit from the stock taking exercise that was done by other WP's in the ESSnet. Until June 2011 the WP was a collaboration between the NSI's of Belgium, the Netherlands and the United Kingdom. In July 2011 Estonia joined the WP.

2. Investigating the Usefulness of Administrative Data

The use of a checklist for evaluating the usefulness of secondary data sources can have a number of advantages:

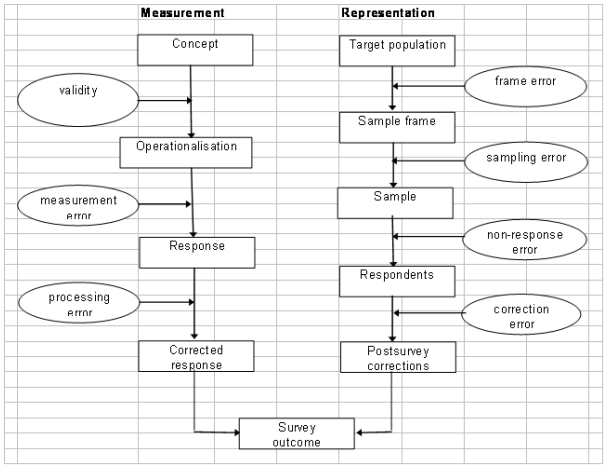
a) It provides a structured way of looking, assuring that the user has paid sufficient attention to the preconditions that are known to be of great importance.

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.
It is in comparing the information noted in the checklist with the requirements for the planned statistical product, that decisions on the usefulness of the data source can be made.

Bart Bakker[2] (Bakker, Linder & van Roon, 2008) develops the general idea that it is likely that the errors that normally emerge in surveys will also occur in administrative data. He refers to Groves et al.[3] (Groves et al., 2004) who describe sources of error (components of total survey error) based on the life cycle of a survey, and concludes that the scheme can easily be adapted to the most common life cycle of a registration. The comparison between checklist elements and categorized "sources of error" can serve to underpin the decision made.



'life cycle' of and errors in a survey (Groves et al., 2004)

3. A More Comprehensive Approach

A checklist for evaluating the usefulness of administrative data reflects the situation at a certain moment in time. Statistical offices wanting to make better use of administrative data have also a clear interest in creating an optimal environment for the use of those data.
WP2 identified three key elements:

- Strengthening cooperation with administrative data holders
- Streamlining transmission and storage of incoming administrative data
- Integrating the pre-processing and first cleaning of administrative data for internal users

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 guidelines and recommendations on the basis of those experiences.

The dynamics of creating the right environment for administrative data would not only lead to more positive outcomes of the checklist, it is also a crucial factor in getting good results when inspecting the incoming stream of administrative data, the topic of WP2b.

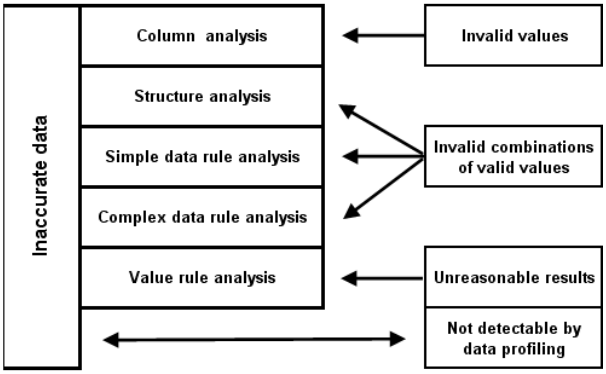
4. Checking the Quality of Administrative Data Inputs

WP2b's objective is to present good practices in checking the incoming administrative data. This should help to reduce the resource needed for the production of business statistics. A reference document will be produced with techniques and guidelines, 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.

4.1 The data-centric approach

The data-centric approach starts from analyzing both data and metadata. The systematic up front analysis of the content of a data source is called data profiling (also referred to as data discovery). It 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. Techniques, such as visualizations, 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.

4.2 The domain expert approach

As an example, some aspects of the work on VAT data are highlighted in this paragraph.

Following a comprehensive literature study, five methods for identifying suspicious VAT Turnover values were tested using two years of UK VAT data, 2004 and 2005, and results were compared.
Once an error or suspicious value has been identified in VAT Turnover data, there are a number of options for dealing with it. The main options are described.
When choosing an option for dealing with errors in VAT Turnover data it is important to consider the statistical uses of the data, the resource available for cleaning the data and the

impact of any possible additional burden on businesses. When the value is changed, it is advisable to keep a record of the original value. This assists with future analysis of the method for changing the value and also ensures that the original data set can be re-created whenever required.
In many cases, the most sensible and cost-effective option will be to automatically change suspicious values using imputation methods. A range of imputation methods are discussed and tested.

5. Conclusions and future work

Administrative data is increasingly being used in the production of statistics as an alternative to or a replacement of survey data.
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. Administrative data on the other hand, are not collected by the statistical institute; the collection process can be very different.
A stocktaking exercise learned that even though administrative data are used widely, and quality issues are considered important, there is no common approach to checking the data.
WP2 shows that there is potential for a comprehensive approach, and has produced a tentative blueprint for more integration of different aspects (like detecting, resolving, and preventing issues).
The results of the WP2 work will be brought together in a reference document, and a first outline of that document has been created. It will contain a checklist, but also recommendations and case descriptions.
It is however only with the advice of the experts in the different Member States that we can build a practical guide with proven techniques, methods, tools and examples of good practices.
Colleagues interested in testing our draft checklist, in duplicating our comparisons of different methods, or who have experience in checking data that they want to share, are welcome to contact the WP.

email: frank.verschaeren@economie.fgov.be

[1] 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.

[2] 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.

[3] Groves, R.M., F.J. Fowler jr., M.P. Couper, J.M. Lepkowski, E. Singer, & R. Tourangeau, 2004, Survey Methodology (New York: Wiley Interscience)