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

Quality control of Dutch Administrative Registers: An inventory of quality aspects

Piet J.H. Daas and Theo C. Fonville
Statistics Netherlands
pdas@cbs.nl, tfne@cbs.nl

1. Introduction

Administrative data are produced as a result of or in connection with the administrative procedures of organizations. Administrative data is becoming an increasingly important data source for the production of statistics by National Statistical Institutes (NSI's) since the use of administrative data drastically reduces the costs and response burden on enterprises and persons. In the Nordic countries administrative registers are the main data source for the production of official statistics [1-4)]. This trend can also be observed in other European countries [5)] and is dictated by:

- Cost reduction. NSI's are confronted with shrinking budgets, and direct data collection (surveys) is expensive;
- Reducing response burden. NSI's need to reduce the response burden on enterprises and persons as much as possible;
- Detailed information requirements. Because administrative data often completely cover whole populations, it is particularly well suited for the creation of detailed information on subpopulations;
- Longitudinal information requirements. Administrative data often cover whole populations over longer periods of time. This enables NSI's to describe the changes over time.

The use of administrative data is further enhanced by the ever increasing use of information and communication technology in public administrations (e-Government) [6)]. As a result of this development, more and more administrative data is becoming available in an electronic form. When public law allows the NSI's to use these electronic administrations [5)], they have the potential of becoming increasingly important data sources for the production of statistics. It is therefore of vital importance that an NSI is able to quickly asses the quality of administrative data - for statistical use - in a standardized way. Reducing assessing time to a minimum requires reliable and efficient methods.

In this report administrative data quality is defined as "*the capability of* [administrative] *data to be used effectively, economically and rapidly to inform and evaluate decisions*" [10)]. First, important aspects of administrative data (i.e., registers) in relation to their statistical use are discussed.

1.1 Registers

Some administrative registers are kept as a basic source for public administration. The function of such registers, called base registers [1)], is typically to keep stock of a specific population of objects [2-4)]. Using the system of registers of the Nordic countries as an example, one can identify three commonly used administrative base registers:

- Population register (persons);

- Property register (real estate, buildings and dwellings);
- Business register (businesses, enterprises and establishments).

In Sweden a fourth base register is defined: the activity register. This register contains information on activities such as jobs, other labour market-related activities and educational activities [2)]. Registers with similar data exist in the other Nordic countries, but they are not defined as base registers [1,3-5)]. Registers that are not considered base registers are denoted as specialized registers [1)]. An example of the latter is the "Student register". This register contains the information of all persons that study at post-primary educational institutes and can, as such, be considered to contain only a subset op the population register. Because of their stock keeping functions, base registers in general contain unique identification codes to identify each of the objects being administrated. Examples of such keys are: personal identification numbers, numerical addresses and business identification numbers. These keys are very important when combining administrative data sources. Fortunately, more and more specialized registers use the same identification keys.

All Nordic NSI's create statistical base registers, which are based on the corresponding administrative registers. The principle tasks of statistical base registers are to define the important populations and contain links to other (statistical) registers. The combined set of statistical base registers forms a register system which is the fundament of the successful use of administrative data in the Nordic NSI's [1-4)].

1.2 Register quality

NSI's have the responsibility to report about their product quality to the users. NSI's of members of the European Community must also report the quality to Eurostat [7,8)]. To do this, standard report sets are available [9)]. For survey-based statistics the creation of such reports is usually fairly straightforward because the entire collection and production process is under the control of the NSI [1-3)]. For statistics predominantly based on administrative register data this is clearly not the case [11)]. Here, the collection and maintenance of the data is beyond the control of the NSI; it is the register holder that manages these aspects. The same is true for the units and variables a register contains. They are defined out of administrative rules and may therefore not be identical to those required by the NSI. Both aspects seriously hamper the determination of the quality of administrative data. This is an important issue as it reduces the (potential) use of administrative data for producing statistics. It is, therefore, of vital importance that NSI's are capable of determining the quality of data in administrative registers - for statistical use - in a quick, straightforward and standardized way. Ideally, evaluation of the quality of an administrative register should result in a report that unequivocally lists the statistical usability of the data. However, until now no widely accepted quality report on the statistical use of registers exists. In addition, no information is given on how any quality indicators of administrative data can be assured objectively with quantitative measures and how they can be used efficiently in practice.

1.3 Quality report

From a practical point of view, two types of reports can be envisaged that describe the quality aspects of administrative data; i) source specific and ii) product specific [9)]. The first type of report discusses the quality issues of the register in general. This is usually the case when the statistical use of the register is unknown in advance. A source specific report describes the quality aspects of a single register. Such a report might therefore contain quality indicators that (to some extent) detect similar quality aspects as those measured by the register holder. This does, however, not have to be the case. The second-type of quality report, the product specific report, will assess the quality indicators of the administrative data's features relevant to the (statistical) product. The latter could very well describe more than one administrative data source; e.g. when two or more registers are used for the creation of a single statistical product [9)]. Some of the quality indicators mentioned in a product specific report might overlap with those in a source specific quality report. Both types of quality reports are relevant for an NSI. In both cases the starting point is the quality aspects of the administrative data source. These quality aspects should be determined, expressed and measured in some way or another. For the determination of the specific quality of register data methods have to be developed that express these aspects in the form of quality indicators. In this report the quality aspects of registers will only be discussed in the source specific way.

1.4 Quality indicators

Eurostat has published a report in which they propose a set of 12 indicators to be used as a measure for the determination of the quality of administrative data [8)]. They differ somewhat from the seven statistical quality components recognized by the same organization [9)]: these indicators are: Relevance, Accessibility and clarity, Completeness, Timeliness and punctuality, Coherence, Comparability and Accuracy [8,9)]. The proposed *administrative* data quality indicators of Eurostat are listed and briefly discussed below:

- Clarity: the result of the evaluation of the metadata documentation of the administrative dataset;
- Administrative concepts: ability to understand the administrative concepts of the data source. The population units, variables and administrative procedures used should be described by the register holder;
- Coverage: the extent of the coverage of the administrative dataset. A precise definition of the population units included in the dataset should be given;
- Reference time: the reference time of the records in the dataset. Is the time recorded the occurrence or the registration of the event or are both recorded? ;
- Data freshness: the time that has lapsed since the last update of the administrative dataset and the likely extent to which the data are outdated;
- Errors in the data: all errors that exist in the data (e.g. measurement, processing and non-response errors). This usually cannot be assessed directly and might imply the assistance of the register holder;
- Completeness: This indicator is mentioned here for completeness. It is only used in product specific reports where it indicates if the administrative data in the register covers all the data needs about the product. In source specific reports this indicator is not used;
- Record matching ability: ability to match the records with those in the (statistical) registers of the NSI's. Any existing common identifiers of population units in the data file should be listed. When this is not the case, the result of the use of other fields for record matching and an evaluation of the effectiveness should be reported;
- Confidentiality and privacy protection: any issues related to confidentiality or privacy protection that may impose constraints on the availability of administrative data to the NSI at a desired level of detail must be reported;
- Compatibility between file formats: comparison between the format in which the administrative data can be made available and the format that can be imported by the NSI. The effect of any conversion efforts should be included;
- Comparability of administrative datasets in time: all necessary information to assess the comparability of the data through time;
- Envisaged use of the data: this item must state what the potential expected use of the data is;

1.5 Approach used

In effect, there are two ways to start the development of a screening method for the determination of the statistical usability of administrative data: a (more) theoretical and a (more) practical approach. For the results described in this paper, the last approach was chosen. The statistical usability of register data studied was leading and the quality indicators, such as those suggested by Eurostat (see above), functioned merely as points of reference. The data files used were:

- Costs of health and social care files from the Central Administration Office ("Centraal Administratie Kantoor Bijzondere Zorgkosten");
- Customer files from Dutch electricity companies ("Energiebedrijven");
- Profit tax files from the Dutch Tax Authorities ("Winstaangifte bestanden van de Belastingdienst").

2. Results

2.1 Preliminary steps

Before the quality of a data source is determined, some things have to be settled in advance:

First, the file has to be identified as being potentially useful. Contact has to be made with the register holder and arrangements on the availability of the data and the way in which it will be delivered to Statistics Netherlands (SN) have to be made. Also privacy considerations must be discussed.

And next, at the start of the evaluation of each data source, a synopsis must be prepared of the register studied. The synopsis contains the name of the register, the abbreviation or short names used (by the register holder and by SN), contact information of the register holder and its contact person(s), the primary contact person at SN for the register holder and any other arrangements made between the register holder and SN. For data sources already in use it can be expected that this information has been documented in some way or another. The synopsis of all data sources that have been evaluated and used by SN should be stored and made centrally available. This does not only enhance the study and use of administrative data sources but will also prevent unnecessary work and eases the subsequent evaluation steps; certainly when the register holder has to be contacted.

2.2 Evaluation of the register

After the creation of a synopsis, the metadata and data of the registers needs to be evaluated subsequently. First the metadata provided by the register holder has to be studied carefully. The metadata must be analyzed first.

2.2.1 Metadata evaluation

The metadata of the register has to be completely and carefully studied. For this the metadata checklist shown in appendix 1 is used. The scores used are explained in appendix 1. The quality aspects used are listed and explained in table 1. After this evaluation step it can be decided if the metadata of the register is described in enough detail to allow further investigation. When each quality aspect scores at least one on availability, actuality and clarity (see appendix 1), this is certainly the case. Any aspect with a lower score should be resolved before the data can be evaluated. If an aspect indicated with an asterisk (*) in table 1 scores less then one and cannot be resolved, the register can certainly not be used by SN. For all other aspects the use of the register is limited to the unaffected data. If the register (or some of the data) can be used the data related aspects can be evaluated.

TABLE 1. Quality aspects of the metadata checklist

Metadata aspects	***Explanation***
Purpose *	What is the original purpose of the registration?
Basis Law / Legal provision / Regulation / Agreements	Legal basis on which the register is kept. Reference to the legal provision or agreement on which the register is based.
Population (conceptual def.) * Geographic demarcation Time demarcation	The population(s) recorded in the register; the object type(s) should be described (e.g. persons, enterprises etc). The geographic area of the population(s) in the register. The period(s) for which the data in the population(s) is registered.
Identification keys *	Unique keys in the register that can be used to identify the recorded object type(s). This could be more than one.
Collection *	The way in which the data is collected by the register holder.
Maintenance *	The way in which the data is maintained by the register holder.
Editing *	If and how the data is edited by the register holder.
Selection	Often SN does not receive a full copy of the register but only a selected set. Check if and what sort of selection is made.
Time dimension * Occurrence Registration	What time event is recorded? Is the time of occurrence recorded for each event. Is the time of registration recorded for each event.
Quality control	Any form of quality control that is (regularly) performed by the register holder.
File format/Data structure *	The file format in which the data is made available.
Classifications / Variable description (key variables *)	Explanation of the classifications and variables used by the register holder.
Supplier agreement *	Agreement between the register holder (data supplier) and SN.
Privacy considerations *	If the register contains unit level identification keys there should be an agreement that the legal rights of the individual citizen with regard to the protection and integrity of his/her data is not violated.

2.2.2 Data evaluation, first stage

The data in the register or the part of the register for which the metadata was described correctly is evaluated for the following quality aspects: coverage and overall reliability. If abnormalities are found during this stage, they should be resolved before more details aspects can be checked.

Coverage

First the coverage of the population(s) of the object types in the register data is determined. The population coverage is checked for completeness. For this a reliable comparison of each unit of each object type(s) in the register should be made with that of the base register(s) for those object type(s). This is no problem for specialized registers but what should be done when a base register itself is being evaluated? At SN, it has been suggested to construct a so-called 'backbone' for each object type from the combined set of all registers that contain the identifier key for the object type [5)]. During the creation of the 'backbone' one can decide to accept all non-erroneous units blindly or one has the difficult decision to decided which units actually belong the population. It is clear that further study is required to decide if such an approach is useful.

The (under-)coverage in a register is expressed as a percentage. The number of non-erroneous units not belonging to the population, the over-coverage, can be calculated in a similar way. The data of the erroneous units and the units not belonging to the population of the object type studied must be ignored in the subsequent analysis. Apart from the coverage issue, the data must also be checked for occurrence of double and erroneous records. Double records can be expected when the register stores the occurrence of each registered event for each unit as a separate record. However, when the time frame of these events overlap, such records are a problem. The register holder should be informed when such errors are found. Erroneous records are records which are certainly wrong. This is obviously the case when an error is found in the check digit of the unique identification key (e.g. social security numbers) of an object type. Occurrence of many errors decreases the reliability and usability of the register data.
In addition to the points mentioned above, the coverage of the population for each object type should also be checked over time. It is important that the units that make up the population can be individually identified through time. It should be clear when units are permanently removed or are newly added to the population. For some registers geographic identification of the objects over time is important as well.

Overall reliability

The data in the register is explored in this stage. Very simple explorative data-analysis (e.g. determination of frequencies, average, medians and totals) should be performed. The results obtained (might) indicate any inconsistencies in the data. Indicators of these are many missing or erroneously coded data; for instance when a lot of data is classified as unknown. The metadata of classifications should correspond with that of the data in the register. An ideal situation occurs when the register can be compared with a previous version of the register (that has already been evaluated). Any quality reports of the register holder should be included in this analysis. Inconsistencies should be reported and discussed with the register holder; although privacy considerations might cause discussion problems here. When large amounts of data are present the exploratory studies should be limited to the set of variables which are -a priori-known to be important for the creation of statistics.

Any coverage problems and unusual data distributions of important variables (for statistics) might prevent further use of the register data. When that is clearly not the case, a more detailed study of the data is required before it definitely can be decided if the register can be used.

2.2.3 Data evaluation, second stage

Data that is found to be correct in the first step of data evaluation needs to be studied in more detail. The following quality aspects need to be further investigated.

Timeliness (data freshness)

The data in the register should describe recent events. The time between the moment that the register is received by SN and the (most recent) period of the events described in the register should not be to long. This, of course, depends largely on the recentness of the statistical topic for which the register is to be used. This is, however, not always known in advance.

Continuity

The register holder should assure SN that the register will be maintained for a certain period in the future. Without such an assurance the register cannot be used. In addition to this, the register holder should inform Statistics Netherland timely of any changes including those resulting from changes in legislation.

Linking

Register data has to be linked with existing data on the micro level. The effectiveness of this should be checked for the identification keys of the object types included in the register. For each object type, the best possible identification key should be used. The percentages of unlinked and erroneous linked records should be determined. Registers which contain multiple records of the same units (resulting in 1:n linkages) and registers which contain non-standard object types (according to SN) may cause problems here. The multiple record problem can be solved by converting the event based records of the register to that of a unit based register. However, a register should only be converted when the study of the other quality aspect reveals that the register is useful for the production of statistics.

Validity

Ideally the data of a selected set of variables should be compared with those of similar data already available at SN. The data should preferably be compared on a more aggregated level because differences at the micro level might not have any effect at the meso or macro level. At this point the data can also be checked for outliers.

Expected use

The combination of the object types and variables present in the register are indicative for the possible statistical use of the data. More specific checks can be performed when (at this stage) it is known for which statistics the data or a selection of the data will be used. When (a part of) the data is already being used for (other) statistics, the problems and solutions found there should be included in the quality report.

3. Conclusions

When the metadata and the data of a register have been completely evaluated, it should be possible to conclude whether a register is useful or a useful addition for the creation of statistics. The evaluation for all metadata quality aspects indicated with an asterisk in table 1 should be found completely correct. For all other metadata aspects a score of less then one (see appendix 1) indicates that that part of the data is useless. The first, global, data analysis should not indicate any coverage problems and unusual data distributions of important variables. Then a second, more detailed, data analysis is performed to obtain information on the usefulness of the data. It is, at this stage, very difficult to quantify a lot of the quality aspects identified. This is a topic that requires further study. The detailed data quality aspects should reveal any limitation in the use of the register data for the production of statistics. After that stage, it should be clear which variables can be used and which not. The part of the register that is found to be of use can then be converted into a statistical register [1,2)].

When the suggested administrative quality indicators of Eurostat are compared with the metadata and data quality aspects discerned by us, its is clear that all quality indicators have been identified in someway or another. We have, especially at the metadata level, broken down the quality indicators in much more (sub)components. In this respect it could be stated that a lot of our quality aspects could be combined into a single quality indicator. From this it is obvious that the Eurostat publication [8)] does indeed propose the important administrative quality components. However, a more detailed study of the relation between the quality aspects we observed and the Eurostat quality indicators is required. In addition to this, two aspects need additionally be considered when evaluating administrative data sources: costs and response burden [1,10)]. Both are a very important stimulant for the utilisation of administrative data in the production of statistics and must therefore be included.

Future work will focus on the creation of a quality report and will include the further evaluation of more registers and the evaluation of registers for which the statistical use is known beforehand. In the near future, a more detailed study on the quantification of the data quality aspects of registers will be performed.

References:

[1] Nordic Statistical Institutes (2007) Register-based statistics in the Nordic countries – Review of the best practices with focus on population and social statistics. UNECE.

[2] Wallgren, A., Wallgren, B. (2007) Register-based Statistics: Administrative Data for Statistical Purposes. Wiley Series in Survey Methodology, United States.

[3] Eurostat, Danmarks Statistik (1995) Statistics on persons in Denmark: a Register-based Statistical System. Theme 0, Series D, Office for Official Publications of the European Communities, Luxembourg, Luxembourg.

[4] Statistics Finland (2004), Use of Register and Administrative Data Sources for Statistical Purposes, Best Practices of Statistics Finland. Handbook 45. Statistics Finland, Helsinki, Finland.

[5] Daas, P, Jeurissen, E., Boonstra, H.J., Nieuwenbroek, N. (2005) Register theory: Registers and Statistics Netherlands (in Dutch) Repport-nr. TMO-R&D-2005-01-31-PDAS Statistics Netherlands, Heerlen, the Netherlands.

[6] Chevallerau, F-X (2005) eGovernment in the Member States of the European Union, main report. IDABC eGovernment Observatory.

[7] Eurostat (2000) Assessment of the quality in statistics, Item4: Definition of quality in statistics. 4-5 April, Luxembourg.

[8] Working group "Assessment of quality in statistics" (2003) Item6: Quality assessments of administrative data for statistical purposes. Sixth meeting, 2-3 October, Luxembourg.

[9] Working group "Assessment of quality in statistics" (2003) Item4.2d: Methodological documents, Handbook "How to make a Quality Report". Sixth meeting, 2-3 October, Luxembourg.

[10] Karr, A.F., Sanil, A.P., Banks, D.L. (2006) Data quality: A statistical perspective. Statist. Methodol. 3, 137-173.

[11] Thomas, M. (2005) Assessing Quality of Administrative Data. Survey Methodol. Bull. 56, 74-84.

Appendix 1.

The following table is used to collect the scores of a set of metadata aspects. Every aspect is scored on the dimensions i) availability, ii) actuality and iii) clarity. For availability, the presence of documentation on each metadata aspect is scored as a 0 (no) or 1 (yes). Documentation that is electronically available is indicated with a plus sign (+). The actuality dimension of the documentation is scored as a 0, 1 or 2; indicating that the documentation is incorrect/incomplete, present but not updated and complete and updated, respectively. Clarity is expressed as 0 or 1; indicating that the documentation is not/very hard or easy/good to interpret, respectively.

Metadata aspects	***Availability*** ***(digital +)***	***Actuality***	***Clarity***
Purpose			
Basis Law / Legal provision / Regulation / Agreements			
Population (conceptual def.) Geographic demarcation Time demarcation			
Identification keys			
Collection			
Maintenance			
Editing			
Selection			
Time dimension Occurrence Registration			
Quality control			
File format/Data structure			
Classifications / Variable description			
Supplier agreement			
Privacy considerations			