Proceedings of Q2008
European Conference on Quality in Official Statistics

Quality framework for the evaluation of administrative data

Piet J.H. Daas[1], Judit Arends-Tóth, Barry Schouten, Léander Kuijvenhoven,
Statistics Netherlands

1. Introduction

National Statistical Institutes (NSI's) collect data for the production of statistics. Apart from the data obtained through surveys, NSI's are increasingly making use of data that is collected and maintained by other organisations for non-statistical purposes. Administrative data is an example of such a data source (Wallgren and Wallgren, 2007). It is produced as a result of administrative processes of organizations but it is -very often- also an interesting data source for NSI's. During the last decade, more and more NSI's have realized this (Unece, 2007). This is especially the case for the NSI's in the Nordic countries. In these countries administrative data is already the main data source for the production of official statistics (Statistics Finland, 2004; Unece, 2007; Wallgren and Wallgren, 2007).

A major advantage of using administrative data for statistics is the fact that it drastically reduces the costs of data collection and the response burden on enterprises and persons. Since administrative data often completely covers whole populations, in various time references, it is also particularly well suited for creating detailed and longitudinal statistics on subpopulations and regions (Wallgren and Wallgren, 2007). An additional stimulus for its use is the increased use of information and communication technology for the exchange of data between (parts of) in public administrations. As a result, more and more administrative data is becoming available in an electronic form that can be easily collected and processed by the NSI's (Børke and Bergstrøm, 2006).

From a statistical point of view, administrative data has some disadvantages as well. The most important one is the fact that the collection and maintenance of administrative data are beyond the control of the NSI. It is the administrator of the data source that manages these aspects. The same is true for the units and variables an administrative data source contains. These are defined by administrative rules and may therefore not be identical to those required by the NSI's (Wallgren and Wallgren, 2007). It often takes considerable effort to clearly determine the statistical usability of administrative data (ESC, 2007; Everaers and Van der Laan, 2003). Since the production of high quality statistics depends on the quality of the input data, it is of vital importance that NSI's are able to determine the quality, i.e. the statistical usability, of administrative data; preferably in a cost efficient way. Although administrative data has been used by statistical offices for quite some time, the determination of the quality of those data sources prior to their use has not received a lot of attention (Unece, 2007; Sæbø et al.,

[1] Corresponding author: Piet Daas, Statistics Netherlands, P.O. Box 4481, 6401 CZ, Heerlen, the Netherlands, pdas@cbs.nl.
The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands

2003). Most of the quality studies performed at NSI's have focused on the quality of data collected by surveys (Biemer and Lyberg, 2003; Van den Brakel et al., 2007) and on the quality of the statistics produced (Eurostat, 2003a-b; Eurostat, 2005b). Only a relatively small number of studies have focused on the quality aspects of administrative data used for statistical purposes (see section 2). Note the word 'aspect' is used in this paper to describe a measurable part of quality.
In this paper an overview is given of the quality framework developed at Statistics Netherlands for the determination of the quality of externally collected data sources. The framework was originally developed for the evaluation of administrative data but early on in the project it was found that it could be applied to other external collected data sources as well. The main goal of the work described in this paper is to identify all quality aspects relevant for the statistical use of external data sources.

2. Statistical quality

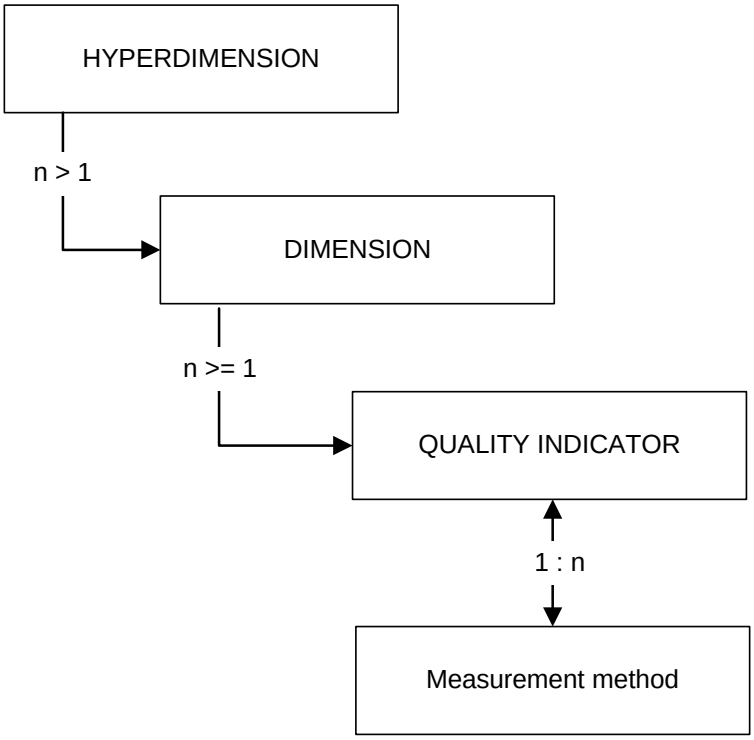
With the adoption of the European Statistics Code of Practice, the NSI's of European Union (EU) member states have committed themselves to an encompassing approach towards high quality statistics (Eurostat, 2005a). NSI's of the EU-member states involved and NSI's of some other European countries, such as Norway, report the quality of their statistical products by using six quality dimensions. The dimensions used are: Relevance, Accuracy, Timeliness and punctuality, Accessibility and clarity, Comparability, and Coherence (Eurostat, 2005b). For the determination of the quality of the input data of NSI's, such as administrative data, the six standard quality dimensions are not always applicable (Eurostat, 2003c). This finding was the starting point for the work described in this paper.

2.1 Quality aspect identification

An extensive literature study revealed that the views on the composition of the quality of administrative data -to be used for statistics- varied greatly. Unfortunately hardly any publications where found that attempted to construct a complete quality framework for administrative data. The most important developments in this area are described in a limited set of papers and books, these are: Wallgren and Wallgren (2007), Daas and Fonville (2007), Eurostat (2003c), Karr et al. (2006), Unece (2007), Thomas (2005), and ONS (2005). When the results of these studies are compared, a clear difference between the number and types of quality groups or dimensions identified for the statistical quality aspects of administrative data is observed. In our opinion this points out the complexity of the problem but also shows that every researcher or group of researchers had a slightly different view on this topic. The progress in this field would be considerable if these heterogeneous views could somehow be combined into a single framework. This exercise was performed by the authors of this paper. The main objective of the work described in this paper was to bring together the different views. The resulting framework should subsequently be able to determine the overall usability of a data source for the production of statistics. Because the statistical usability is determined at a general level, such a framework will only consist of quality indicators that can be generally applied (see also section 2.4).
By combining the various quality aspects identified at Statistics Netherlands (Daas and Fonville, 2007) and those mentioned in the publications of others (listed above), the authors attempted to get a complete overview of all the quality aspects of administrative

data relevant for statistical use. Every quality aspect identified in every study was compared with those observed in any of the other studies. During this exercise two important findings emerged. The first one is the fact that -despite of the differences- there is a general level of mutuality. In a lot of studies many (very) similar quality aspects were identified. The second one is the observation that the statistical quality of administrative data is more than a simple dimensional concept. Depending on the perspective from which the administrative data source is looked upon, different quality aspects prevail. Such a perspective, a high level *view* at statistical quality, has been described by Karr et al. (2006). In their point of view, statistical quality not only consists of dimensions and indicators but also of a concept they identified as a 'hyperdimension'. A hyperdimension is a way of looking at quality at a level higher than that of the commonly used dimension; hence the name 'hyper'dimension.
In a hyperdimension several dimensions of quality are grouped. The quality aspects included are highly influenced by the contextual view on the quality of the data source (Karr et al. 2006). With the above in mind, a quality framework was developed for administrative data that consisted of hyperdimensions, dimensions, quality indicators, and measurement methods (figure 1). A hyperdimension is composed of two or more dimensions and each dimension contains one or more quality indicators. A quality indicator is measured or estimated by one or more methods. The relation between the various quality aspects included in the framework is shown in figure 1. With the identified hyperdimensions in mind it became fairly easy to allocate every quality aspects found and group them into dimensions.

Figure 1. Hierarchical relation between the different aspects of quality used in the framework developed for administrative data



2.2 Quality framework

The identification and comparison of all the quality aspects identified for administrative data revealed four discernible contextual ways of looking at the quality of such a data source. The four hyperdimensions identified were called: Source, Metadata, Data, and Process. The quality indicators in the first three hyperdimensions (Source, Metadata, and Data) all have a product based approach in common (Ehling and Körner, 2007; Eurostat, 2002). The indicators in these three dimensions each highlight different quality aspects of the data source. The Source, Metadata, and Data hyperdimensions are also ordered according to an increase in the level of detail. The quality indicators in the Data hyperdimension, for instance, report on quality aspects in the data source at a much more detailed level than the quality indicators included in the Metadata hyperdimension. The same is true for the Metadata and Source hyperdimensions. The Process hyperdimension specifically focuses on the underlying process by which the data source is created and used. The process starts with the collection of data by the data source maintainer and ends with the receipt of the data source by the NSI. In this paper only the product-based hyperdimensions, viz. Source, Metadata, and Data, are discussed. In tables 1-3 an overview is given of the dimensions, quality indicators, and measurement methods for these hyperdimensions.

In the Source hyperdimension the data source is considered a file that is delivered by the data source maintainer to the NSI. In table 1 the dimensions, quality indicators, and measurement methods for the Source hyperdimension are listed. In this hyperdimension mainly qualitative methods are present. An exception is the calculation of the effect of the use of the data source on the administrative burden of the NSI.

The Metadata hyperdimension specifically focuses on the metadata aspects of the data source. In table 2 the dimensions, quality indicators, and measurement methods are listed for the Metadata hyperdimension. The Metadata hyperdimension solely contains qualitative methods. The Data treatment dimension of the Metadata hyperdimension is a special case. It consists of quality indicators that are used to determine if the data source maintainer performs any checks and/or modifications of the data in the source. Although this information can be considered as meta-information of the process (and should as such be included in the Process hyperdimension) it is also very important for the final quality of the administrative data source. Because of this, the Data treatment dimension was included in the Metadata hyperdimension.

The Data hyperdimension focuses on the quality aspects of the data in the data source. This hyperdimension predominantly contains accuracy related quality aspects (table 3) with the exception of the Technical Checks dimension. The latter dimension contains indicators that verify the readability of the data file and the compliance of the data to the metadata definition. A considerable part of the measurement methods in the Data hyperdimension are based on the so-called Representativity index (R-index). An R-index is a concept that has been developed by Statistics Netherlands (Schouten and Cobben, 2007; Cobben and Schouten, 2008). R-indices measure the extent to which the composition of the units in a data source, at a certain point in time, deviate from the population. For surveys this is a familiar concept. Here, representative means that all units in the population have the same probability of responding to the survey request (Cobben and Schouten, 2008; Ariel and Schouten, 2008). Representative is, however, also important for administrative data because the composition of the units present in the data source may be time-dependent. In the Netherlands, for instance, the

Table 1: Quality framework for data sources, Source hyperdimension

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Supplier	1.1 Contact	- Name of the data source - Data source contact information - NSI contact person
	1.2 Purpose	- Reason for use of the data source by NSI
2. Relevance	2.1 Usefulness	- Importance of data source for NSI
	2.2 Envisaged use	- Potential statistical use of data source
	2.3 Information demand	- Does the data source satisfy information demand?
	2.4 Response burden	- Effect of data source on response burden
3. Privacy and security	3.1 Legal provision	- Basis for existence of data source
	3.2 Confidentiality	- Does the Personal Data Protection Act apply? - Has use of the data source been reported by NSI?
	3.3 Security	- Manner in which the data source is send to NSI - Are security measures required? (hard-/ software)
4. Delivery	4.1 Costs	- Costs of using the data source
	4.2 Arrangements	- Are the terms of delivery documented? - Frequency of deliveries
	4.3 Punctuality	- How punctual can the data source be delivered? - Rate at which exceptions are reported - Rate at which data is stored by data source maintainer
	4.4 Format	- Formats in which the data can be delivered
	4.5 Selection	- What data can be delivered? - Does this comply with the requirements of NSI?
5. Procedures	5.1 Data collection	- Familiarity with the way the data is collected
	5.2 Planned changes	- Familiarity with planned changes of data source - Ways to communicate changes to NSI
	5.3 Feedback	- May NSI contact data source maintainer in case of trouble? - In which cases and why?
	5.4 Fall-back scenario	- Dependency risk of NSI - Emergency measures when data source is not delivered according to arrangements made

composition of the companies that provide VAT-data to the Dutch Tax Office varies during the monthly collection period. This affects the quality of the data that is provided to Statistics Netherlands. Because of the fact that time-related data quality issues are included in R-indices, timeliness is not added as a separate dimension in the Data hyperdimension. As can be seen in table 3, the dimension Precision is also included in the Data hyperdimension. For administrative data, this dimension is mainly used to determine the effect of time-dependent changes in the population composition on data quality (see also section 2.3).

As mentioned before, it is important to realize that the quality framework was created to be used at a general level. The framework therefore only contains quality indicators that can be generally applied. Very specific checks are not included, simply because it is

Table 2: Quality framework for data sources, Metadata hyperdimension

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Clarity	1.1 Population variable definition	- Clarity score of the definition
	1.2 Classification variable definition	- Clarity score of the definition
	1.3 Count variable definition	- Clarity score of the definition
	1.4 Time dimensions	- Clarity score of the definition
	1.5 Definition changes	- Familiarity with occurred changes
2. Comparability	2.1 Population variable definition comparison	- Comparability with NSI definition
	2.2 Classification variable definition comparison	- Comparability with NSI definition
	2.3 Count variable definition comparison	- Comparability with NSI definition
	2.4 Time differences	- Comparability with NSI reporting periods
3. Unique keys	3.1 Identification keys present	- Presence of unique keys - Comparability with unique keys used by NSI
	3.2 Unique combinations of variables present	- Presence of useful combinations of variables
4. Data treatment (by data source maintainer)		
	4.1 Checks	- Population unit checks performed - Variable checks performed - Combinations of variables checked - Extreme value checks
	4.2 Modifications	- Familiarity with data modifications - Are modified values marked and how? - Familiarity with default values used

impossible to include all possible conceivable checks. An example of a specific check is the comparison between the results obtained for ‘the percentage of unemployment persons in the Netherlands’ (for a specific month) derived from an administrative data source and from a survey of Statistics Netherlands (see section 2.4). It’s clear that such a check should not be included in the framework because it can’t be generally applied. Another reason for only including general applicable indicators is the fact that different users of a data source may have different population parameters in mind that pose different quality constraints. Necessarily, the quality framework has to be restricted to some extent as it is impossible to meet all conceivable uses.

2.3 Quality framework and external data sources

The content of the Source, Metadata, and Data hyperdimension of the quality framework constructed for administrative data sources is shown in tables 1 through 3. Although the quality framework was originally developed for administrative data, it also interested the authors to see if it could be used for the evaluation of other externally collected data sources as well. One of such sources is survey data collected by an organization other than Statistics Netherlands. Initial evaluation results indicated that the framework could indeed be applied to externally collected survey data. Only the wordings of some of the

Table 3: Quality framework for data sources, Data hyperdimension

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Technical checks	1.1 Readability 1.2 Metadata compliance	- Can all the data in the source be accessed? - Does the data comply to the metadata definition? - If not, report the anomalies
2. Over coverage	2.1 Non-population units	- Percentage of units not belonging to population - Percentage of duplicate units
3. Under coverage	3.1 Missing units	- Percentage of missing population units
	3.2 Selectivity	- R-index [a] for population composition
	3.3 Effect on core variables	- Maximum bias of average for core variable - Maximum RMSE [b] of average for core variable
4. Linkability	4.1 Linkable units	- Percentage of units linked
	4.2 Mismatches	- Percentage of units incorrectly linked
	4.3 Selectivity	- R-index for units linked
	4.4 Effect on core variables	- Maximum bias of average for core variable - Maximum RMSE of average for core variable
5. Unit non response	5.1 Units without data	- Percentage of units without data
	5.2 Selectivity	- R-index for unit composition
	5.3 Effect on core variables	- Maximum bias of average for core variable - Maximum RMSE of average for core variable
6. Item non response	6.1 Missing values	- Percentage of cells with missing values
	6.2 Selectivity	- R-index for variable composition
	6.3 Effect on variable	- Maximum bias of average for variable - Maximum RMSE of average for variable
7. Measurement	7.1 External check	- Has an audit or parallel test been performed? - Has the input procedure been tested?
	7.2 Incompatible records	- Fraction of fields with violated edit rules
	7.3 Measurement error	- Size of the relative measurement error
8. Processing	8.1 Adjustments	- Fraction of fields adjusted
	8.2 Imputation	- Fraction of fields imputed
	8.3 Outliers	- Fraction of fields with outliers
9. Precision	9.1 Standard error	- MSE [c] of average value for core variable
10. Sensitivity	10.1 Missing values	- Total percentage of empty cells
	10.2 Selectivity	- R-index for composition of totals
	10.3 Effect on totals	- Maximum bias of average for totals - Maximum RMSE of average for totals

[a] R-index, Representative index (explained in section 2.2.3); [b] RMSE, Root Mean Square Error; [c] MSE, Mean Square Error.

measurement methods in the three hyperdimensions had to be adjusted to enable its -more general- use to both (externally collected) survey and administrative data sources. These wordings are those shown in table 1-3. For survey data, the inclusion of the Precision dimension is essential (table 3). There are additional advantages of extending the quality framework to that of other externally collected data sources. The most important one is the fact that this enables the use of a single framework for the determination of the quality of (possibly) all externally collected data sources used for the production of statistics. Currently more detailed evaluation studies are performed to verify this initial finding.

2.4 Use of the quality framework

The framework presented in this paper can be used for the evaluation of the quality of a data source (a product). The quality of a product should be determined by successively evaluating the quality aspects of the Source, Metadata, and Data hyperdimensions. This strict order is the result of the fact that the quality aspects in the Source hyperdimension report on quality at a much more general level than the aspects included in the Metadata and Data hyperdimensions. The same is true for the quality aspects of the Metadata hyperdimension in comparison to those of the Data hyperdimension (see section 2.2).
As a consequence, the user must first evaluate the quality indicators in the Source hyperdimension, then those in the Metadata hyperdimension, and finally those in the Data hyperdimension. When the results for some of the quality indicators in a hyperdimension reveal problems, these have to be sorted out before the next hyperdimension can be evaluated. If these problems cannot be solved, the evaluation of the data source must be stopped and it has to be concluded that the data source cannot be used for the statistical application the user had in mind. When unsolvable problems occur during the evaluation of the Source hyperdimension it is likely that the user has to conclude that the data source can not be used for statistics at all. When the user has another (new) statistical use in mind for a data source that has already been evaluated, the same sequence of events must be repeated. It is very likely, however, that results for the Source hyperdimension won’t differ a lot from those previously obtained.
For the evaluation of the Source and Metadata hyperdimension, the authors have developed checklists. These guide the user through the sets of quality indicators that need to be evaluated in each hyperdimension. When problems are found or a question can not be answered completely, the user is also guided in the steps to take. The Source and Metadata checklists can be used for a data source that is already available to the NSI and for the evaluation of a new data source that could potentially be used for statistics. For the Data hyperdimension a checklist can not be used because of the large amount of calculations that need to be performed. How the quality indicators in this hyperdimension should routinely be evaluated is a topic of current research.
If the evaluation of the last hyperdimension, Data, is successful, the data source can be used for the production of statistics. It is conceivable, however, that the user would like to perform one or more additional -very specific- checks after the evaluation of the three hyperdimensions. These additional checks will occur at the data level. An example of such a specific check is a comparison of the estimated percentage of unemployed persons obtained, after editing and weighting, from an administrative data source (such as the Job-seeker administration of the Centre of Work and Income in the Netherlands) with that of the estimated percentage obtained through the Labour Force survey of

Statistics Netherlands. These types of checks are not included in the quality framework because, as stated before (sections 2.1 and 2.2), the framework only contains general applicable quality indicators.

2.5 Future work

Future studies will focus on the usability of the quality framework on both administrative and survey data sources. In these studies the checklists constructed for the Source and Metadata hyperdimensions will be used and evaluated. For the Data hyperdimension another approach is followed. The development of a set of standards scripts or inclusion of the measurement methods in a specific software program are two of the options currently explored.

References

Ariel, A., and B. Schouten (2008), “Representativity of the time use survey”, Discussionpaper 08013, Voorburg/Heerlen, The Netherlands: Statistics Netherlands.

Biemer, P. P., and L. E. Lyberg (2003), *Introduction to Survey Quality*, New Jersey: Wiley.

Børke, S., and Y. Bergstrøm (2006), “Electronic Data from Business to Government - Development with Implications on Administrative Registers?”, paper presented at the seminar on Strategies for Social and Spatial Statistics, Oslo, Norway.

Cobben, F., and B. Schouten (2008), “An empirical validation of R-indicators”, Discussionpaper 08006, Voorburg/Heerlen, The Netherlands: Statistics Netherlands.

Daas, P. J. H., and T.C. Fonville (2007), “Quality control of Dutch Administrative Registers: An inventory of quality aspects”, paper presented at the seminar on Registers in Statistics - methodology and quality, Helsinki, Finland.

ESC (2007), “Pros and cons for using administrative records in statistical bureaus”, paper presented at the seminar on increasing the efficiency and productivity of statistical offices, Economic and Social Council conference of European statisticians, Geneva, Switserland.

Ehling, M., and T. Körner (2007), *Handbook on Data Quality Assessment Methods and Tools*, Wiesbaden: European Commission, Eurostat.

Eurostat (2002), “Quality in the European Statistical System – The Way Forward”, Luxembourg, Luxembourg: Office for Official Publications of the European Communities.

Eurostat (2003a), “Item 4.2 Methodological documents Definition of quality in statistics”, Luxembourg: Working group Assessment of quality in statistics, Eurostat.

Eurostat (2003b), “Item 4.2d Methodological documents Handbook How to make a Quality Report”, Luxembourg: Working group Assessment of quality in statistics, Eurostat.

Eurostat (2003c), “Item 6 Quality assessments of administrative data for statistical purposes”, Luxembourg: Working group Assessment of quality in statistics, Eurostat.

Eurostat (2005a), *European Statistics Code of Practice for the national and community statistical authorities*, Luxembourg: Eurostat.

Eurostat (2005b), “Standard quality indicators”, Luxembourg: Taskforce quality in statistics, Eurostat.

Everaers, P. C. J., and P. van der Laan (2003), “The Dutch System of Social Statistics: Micro-Integration of Different Sources”, paper presented at the expert group meeting on Setting the Scope of Social Statistics, United Nations, New York, USA.

Karr, A. F., A. P. Sanil, and D. L. Banks (2006), “Data quality: A statistical perspective”, *Statistical Methodology*, 3, pp. 137-173.

ONS (2005), “Guidelines for measuring statistical quality”, version 3.0, London: Office of National Statistics.

Sæbø, H. V., J. Byfuglien, and R. Johannesen (2003), “Quality Issues at Statistics Norway”, *Journal of Official Statistics*, 19, pp. 287-303.

Schouten, B., and F. Cobben (2007), “R-indexes for the comparison of different fieldwork strategies and data collection modes”, Discussionpaper 07002, Voorburg/Heerlen, The Netherlands: Statistics Netherlands.

Statistics Finland (2004), *Use of Register and Administrative Data Sources for Statistical Purposes*, Handbook 45, Helsinki: Statistics Finland.

Thomas, M. (2005), “Assessing Quality of Administrative Data”, *Survey Methodological Bulletin*, 56, pp. 74-84.

Unece (2007) *Register-based statistics in the Nordic countries – Review of best practices with focus on population and social statistics*, Geneva: United Nations Publication.

Van den Brakel, J., P. Smith P., and S. Compton (2007), “Quality procedures for survey transitions, experiments and discontinuities”, Discussion paper 07005, Voorburg/Heerlen, The Netherlands: Statistics Netherlands.

Wallgren, A., and B. Wallgren (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*, Chichester: Wiley.