Proposal for a quality framework for the evaluation of administrative and survey data*

Piet J.H. Daas#

Judit Arends-Tóth

Barry Schouten

Léander Kuijvenhoven

Statistics Netherlands, Kloosterweg 1, 6412 CN Heerlen, The Netherlands,
e-mail: pdas@cbs.nl

Abstract: Statistics Netherlands is increasingly making use of data sources collected and maintained by others, such as administrative data, for the production of statistics. Since the quality of the statistics produced is affected by the quality of the data sources used, it is of vital importance that Statistics Netherlands is able to unambiguously determine the quality, i.e. the statistical usability, of external data sources. For this purpose a quality framework was developed for administrative data. The framework is discussed in this paper. It was found that the framework could also be used for the evaluation of survey data. As such, a single framework could potentially be used for the evaluation of all input data sources used for the production of statistics.

Keywords: Quality framework, Administrative Data, Survey Data, Quality aspects.

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 others 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 the use of administrative data for statistics is the fact that it drastically reduces the costs of data collection and the response burden on enterprises

---

* This paper was written for the workshop on the "Combination of surveys and administrative data" that was held on 29-30th May 2008 in Vienna, Austria. The workshop was organized by the ESS-net "Integration of survey and administrative data". BPA number: DMH-2008-04-23-PDAS.

The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

and persons. Since administrative data often completely covers whole populations, in various time references, it is also particularly well suited for the creation of 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 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 (Børke and Bergstrøm 2006).

From a statistical point of view, administrative data has some disadvantages. 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 out of administrative rules and may therefore not be identical to those required by the NSI (Wallgren and Wallgren 2007). It often takes considerable effort to unambiguously 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 unambiguously 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 there use has not received a lot of attention (UNECE 2007, Sæbø et al. 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, 2005b). Only a relative small number of studies has focused on the quality aspects of administrative data used for statistical purposes (see section 2). For all clarity, 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 for the determination of the quality of administrative and survey data at Statistics Netherlands. The framework was originally developed for the evaluation of administrative data but early on in the project it was found that it could also be applied to survey data. The main goal of the work described in this paper is to identify all quality aspects relevant for the statistical use of data sources.

2. Statistical quality

With the adoption of the European Statistics Code of Practice, the NSI's of 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. This was also highlighted in a publication of Eurostat (Eurostat 2003c). The study of the quality aspects of administrative data 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 remarkable 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 suggests 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 this paper is to bring together these different views.

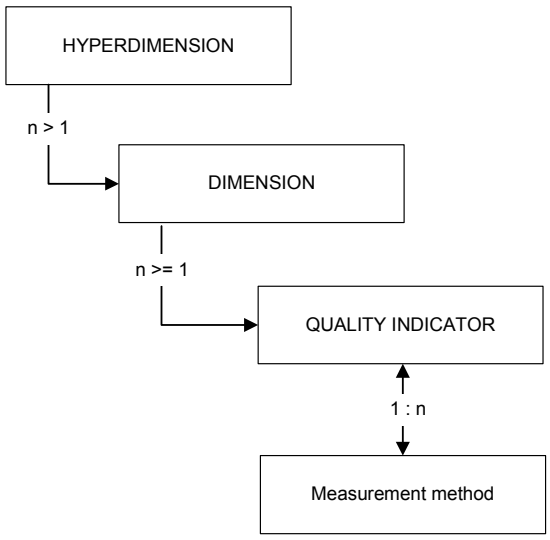
By combining the various quality aspects identified at Statistics Netherlands (Daas and Fonville 2007) and those mentioned in the publications of others (listed above), the authors attempted to get a complete overview of the quality aspects of administrative data relevant for statistical use. Every quality aspect identified in every study was compared with those observed in any of the other studies. During this exercise two important findings emerged. First, there is a general level of mutuality; in a lot of studies many (very) similar quality aspects were identified. Second, 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, was described by Karr et al. (2006). In there 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 a 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 consists 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.

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. Each hyperdimension highlights specific quality aspects of the data source. The quality indicators in the first three hyperdimensions (Source, Metadata, and Data) are all different. These three hyperdimensions are also ordered according to an increase in the level of detail. The quality indicators in the Data hyperdimension, for instance, report on

Figure 1: *Relation between the different aspects of quality in the framework developed*



quality aspects in the data source at a much more detailed level than the quality indicators included in the Metadata hyperdimension. This same is true for the Metadata and Source hyperdimensions. The four hyperdimensions identified are briefly discussed below. More detailed information is presented in tables 1, 2, and 3.

2.2.1 Source hyperdimension

In the Source hyperdimension the data source is viewed upon as a file that is delivered by the data source maintainer to the NSI. Only quality aspects related to this view are included in this hyperdimension. In the Source hyperdimension, five dimensions are distinguished: Supplier, Relevance, Privacy and security, Delivery, and Procedures (table 1). Examples of quality indicators in the Source hyperdimension are: contact information of the NSI, effect on response burden, costs of data source use, data source delivery arrangements, communication of planned changes and dependency risks. Mainly qualitative methods are present in this hyperdimension, only a few quantitative methods occur. In table 1 the dimensions, quality indicators, and measurement methods for the Source hyperdimension are listed.

2.2.2 Metadata hyperdimension

The Metadata hyperdimension specifically focuses on the conceptual metadata of the data source. Clarity of the definitions and completeness of the meta information are some of the quality aspects included. The Metadata hyperdimension is composed of four dimensions: Clarity, Comparability, Unique keys, and Data treatment by data source maintainer. Examples of quality indicators in the Metadata hyperdimension are: clarity of the population definition, time differences between the reporting periods of the NSI and data source maintainer, presence of unique identification keys, and data checks performed by the data source maintainer. The Metadata hyperdimension solely contains qualitative methods. In table 2 the dimensions, quality indicators, and measurement methods are listed for the Metadata hyperdimension.

2.2.3 Data hyperdimension

The Data hyperdimension focuses on the quality aspects of the data in the data source. This hyperdimension solely contains accuracy related quality aspects. The Data hyperdimension is composed of nine dimensions: Over coverage, Under coverage, Linkability, Unit non-response, Item non-response, Measurement, Processing, Precision, and Sensitivity. The dimension Precision was added mainly because of its application for surveys (see section 2.3). Examples of quality indicators are: units not belonging to the population, correctly coupled units, missing values, and measurement error. The Data hyperdimension almost exclusively contains quantitative methods. In table 3 the dimensions, quality indicators, and measurement methods for the Data hyperdimension are listed.
A considerable part of the measurement methods in the Data hyperdimension are based on a so-called Representative index (R-index). The R-index is a concept that has been developed by Statistics Netherlands (Schouten and Cobben 2007). R-indexes 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. 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 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.

2.2.4 Process hyperdimension

The Process hyperdimension is quite different in comparison to the other three. In the Process hyperdimension the focus is not on the data source itself but on the processing of the data source. Therefore, it was excluded in our initially studies and is not discussed in this paper. It is, however, a subject that is under investigation in our office.

2.3 Quality framework and survey data

The overall quality framework constructed is shown in tables 1 through 3. For each hyperdimension a separate table is created that shows its dimensions, quality indicators, and methods of determination. 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 survey data as well. The initial reason for doing this was to see if the framework could be used to determine the quality of survey data collected by an organization other than Statistics Netherlands. Evaluation results indicated that this could indeed be the case. To enable the use for survey data it was, however, required to add some survey specific quality indicators; such as those included in the Precision dimension (table 3). The time-dependence of the population composition in administrative data is another, non-survey specific, reason for doing this (section 2.2.3). In the Source and Metadata hyperdimension only the wordings of some of the measurement methods had to be adjusted to enable its use for survey and administrative data sources.

There are additional advantages of extending the administrative data quality framework to that of surveys. The most important one is the fact that this enables the use of a single framework for the determination of the quality of the two most important data sources used for the production of statistics in our office. Currently more detailed evaluation studies are performed to verify this initial finding.

2.4 Use of the quality framework

While evaluating a (potential) data source, the user of the quality framework must first determine the results for the quality indicators in the Source hyperdimension. Subsequently, the quality indicators in the Metadata and Data hyperdimension need to be evaluated. The authors have developed a checklist for the evaluation of the first two hyperdimensions. This approach could not be followed for 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 be concluded that the data source cannot be used for the statistics the user had in mind. If the user wants to evaluate the data source again but with another (new) statistical use in mind, the same sequence of events should be repeated. However, if the problems for that data source occurred in the Source hyperdimension it is to be expected that the data source can also not be used.
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 all hyperdimensions. The additional checks all occur at the data level; i.e. in the Data hyperdimension. 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. Since the quality framework was created to be used at a general level, it only contains general applicable quality indicators. Very specific check were not included, simply because it is impossible to include all possible conceivable checks. 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. The necessity to restriction is especially applicable to the quality indicators of the Data hyperdimension.

2.5 Future work

Future studies will focus on the usability of the quality framework on both administrative and survey data. In these studies the checklists constructed for the Source and Metadata hyperdimensions will be evaluated. For the Data hyperdimension another approach will be followed; one of the options is to include the measurement methods in a specific software program. Various registers and surveys on persons and businesses will be evaluated. The study of the quality aspects of the Process hyperdimension is also a topic for future work.

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 [a] contact person
	1.2 Purpose	- Reason for use of the data source by NSI
2. Relevance		
	2.1 Usefulness	- Importance 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- and 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 in which changes are communicated to NSI
	5.3 Feedback	- May NSI contact data source maintainer is 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

[a)] NSI, National Statistical Institute

Table 2: *Quality framework for data sources, Metadata hyperdimension*

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

Table 3: *Quality framework for data sources, Data hyperdimension*

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Over coverage	1.1 Non-population units	- Percentage of units not belonging to population
2. Under coverage	2.1 Missing units	- Percentage of missing population units
	2.2 Selectivity	- R-index [a] for population composition
	2.3 Effect on core variables	- Maximum bias of average for core variable - Maximum RMSE [b] of average for core variable
3. Linkability	3.1 Linkable units	- Percentage of units linked
	3.2 Mismatches	- Percentage of units incorrectly linked
	3.3 Selectivity	- R-index for units linked
	3.4 Effect on core variables	- Maximum bias of average for core variable - Maximum RMSE of average for core variable
4. Unit non response	4.1 Units without data	- Percentage of units without data
	4.2 Selectivity	- R-index for unit composition
	4.3 Effect on core variables	- Maximum bias of average for core variable - Maximum RMSE of average for core variable
5. Item non response	5.1 Missing values	- Percentage of cells with missing values
	5.2 Selectivity	- R-index for variable composition
	5.3 Effect on variable	- Maximum bias of average for variable - Maximum RMSE of average for variable
6. Measurement	6.1 External check	- Has an audit or parallel test been performed? - Has the input procedure, e.g. questionnaire, been tested?
	6.2 Incompatible records	- Fraction of fields with violated edit rules
	6.3 Measurement error	- Size of the relative measurement error
7. Processing	7.1 Adjustments	- Fraction of fields adjusted
	7.2 Imputation	- Fraction of fields imputed
	7.3 Outliers	- Fraction of fields with outliers
8. Precision	8.1 Standard error	- MSE [c] of average value for core variable
9. Sensitivity	9.1 Missing values	- Total percentage of empty cells
	9.2 Selectivity	- R-index for composition of totals
	9.3 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

References

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

Børke S., Bergstrøm Y. (2006) Electronic Data from Business to Government - Development with Implications on Administrative Registers?, *Seminar on Strategies for Social and Spatial Statistics*, Oslo.

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.

ESC (2007) Pros and cons for using administrative records in statistical bureaus, *Seminar on increasing the efficiency and productivity of statistical offices*, Economic and Social Council conference of European statisticians, Geneva.

Eurostat (2003a) *Definition of quality in statistics*, Assessment of the quality in statistics, Item 4.2: Methodological documents, Luxembourg.

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

Eurostat (2003c) *Quality assessments of administrative data for statistical purposes*, Assessment of quality in statistics, Item 6, Luxembourg.

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

Eurostat (2005b) *Standard quality indicators*, Quality in statistics, Luxembourg.

Everaers P. C. J., Van der Laan P. (2003) The Dutch System of Social Statistics: Micro-Integration of Different Sources, *Expert Group Meeting on Setting the Scope of Social Statistics*, United Nations, New-York.

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

ONS (2005) *Guidelines for measuring statistical quality*, version 3.0, Office of National Statistics, United Kingdom.

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

Schouten B., Cobben F. (2007) *R-indexes for the comparison of different fieldwork strategies and data collection modes*. Discussion paper 07002, Statistics Netherlands, Voorburg/Heerlen.

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

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

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

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

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