Overall Assessment of the Quality of Administrative Data Sources

Ossen, Saskia J.L.*
Statistics Netherlands, Division of Methodology and Quality
CBS-weg 11
6412 EX, Heerlen, The Netherlands
E-mail: sjl.ossen@cbs.nl

Daas, Piet J.H.
Statistics Netherlands, Division of Methodology and Quality
CBS-weg 11
6412 EX, Heerlen, The Netherlands
E-mail: pjh.daas@cbs.nl

Tennekes, Martijn
Statistics Netherlands, Division of Methodology and Quality
CBS-weg 11
6412 EX, Heerlen, The Netherlands
E-mail: m.tennekes@cbs.nl

Abstract: The quality of administrative data sources has been studied quite intensively in the Netherlands since a number of years. The development of a quality framework that is able to determine the statistical 'usability' of administrative data sources is the major product of this study. Focus of the framework is the evaluation of the quality of an administrative data source in the beginning of the statistical process. The framework developed consists of three high level views on the quality of administrative sources. With a checklist the quality aspects included in the first two views, called Source and Metadata, can be determined. The first view predominantly focuses on the exchange of the data source with the data source holder, while the second view focuses on the metadata of the data in the source. Major advantage of the checklist is that the quality components in the two first views can be determined independently of the data in the source and, as a result, do not have to be checked every time a new dataset arrives. The study of the quality of the data, the third view in the framework, is not included in the checklist and is the topic of current research. Up till now quality indicators have been identified for administrative data when used as an input source for the statistical process. These indicators have been grouped according to the following five general dimensions of data quality: Technical checks, Accuracy, Completeness, Integrability, and a Time-related dimension. If applicable, a distinction is made in each dimension between quality indicators specific for objects (such as units and events) and for variables. Currently measurement methods for these indicators are developed and tested. This paper gives an overview of the framework and discusses its use for the quality evaluation of administrative sources.

Key words and phrases: Register quality, Administrative data, Register-based statistics.

1. Introduction

National Statistical Institutes (NSI's) need data for the production of statistics. Apart from data obtained through surveys, NSI's are increasingly using data collected and maintained by other, non-statistical,

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

organizations. 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; Frost et al., 2010). A major advantage of using administrative data for statistics compared to survey data is that it reduces the costs of data collection and reduces the administrative burden on enterprises and persons. Since administrative data often cover whole populations, it is also very well suited for creating detailed and longitudinal statistics on subpopulations and regions (Wallgren and Wallgren, 2007).

A generic problem in using administrative data sources for statistical purposes is that these sources are collected and maintained by other organizations for non-statistical purposes. This is a process that is beyond the control of NSI's. These institutes highly depend on this process for it may affect the quality of the output. Since many NSI's are expected to use more and more administrative data in the future, a quality framework has been developed that enables the determination of the quality of externally collected data sources, such as administrative data sources and registers, prior to use. The framework focuses on the *input* side of the statistical process (Daas et al., 2010; 2011) and contains all quality components identified for statistical use. Readers are referred to the papers of Daas et al. (2008a-b, 2009, 2010, 2011) for a more detailed overview of the development of the framework and the choices made. The present paper gives an overview of the final version of the framework and describes how it should be applied.

2. Quality framework

The quality framework for administrative data sources was developed to standardize the determination of the various quality components of those data sources (Daas et al., 2008b) looked upon from the *input* site of the statistical process (Daas et al., 2010). At the highest level, the quality framework consists of three views on quality. These views are referred to as hyperdimensions and are called: Source, Metadata, and Data (Daas et al., 2009). Each hyperdimension is composed of several dimensions of quality and each dimension contains a number of quality indicators. A quality indicator is measured or estimated by one or more methods which can be qualitative or quantitative. Subsection 2.1 starts with an overview of the quality components in the Source and Metadata hyperdimension and the methods developed to determine them. Next, recent insights on the study of the quality components in the Data hyperdimension are described (Daas et al., 2011).

2.1 Source and Metadata hyperdimensions

An NSI that plans to use an administrative data source should start by exploring the quality of the arrangements that enable the use of the data source on a regular basis. These components of quality are located in the Source hyperdimension of the framework. In table 1 the dimensions, quality indicators, and method descriptions for this hyperdimension are shown. The second hyperdimension in the framework, called Metadata, focuses on the conceptual and process related quality components of the metadata of the source. It is essential that an NSI, prior to use of the data source, has fully reviewed the metadata related quality components listed in the framework. In table 2 the dimensions, quality indicators, and method descriptions are shown for the Metadata hyperdimension. For the evaluation of the quality indicators in the Source and Metadata hyperdimension a checklist has been developed. It is included in the 2009-discussion paper of Daas et al. (2009).

2.2 Data hyperdimension

Indicators for the quality of the data are part of the Data hyperdimension. These indicators focus on the quality of the data used as input in the statistical process (Daas et al., 2010; 2011). The indicators are grouped according to five dimensions: Technical checks, Accuracy, Completeness, Time-related, and Integrability (Daas et al., 2011). If applicable, a distinction is made between quality indicators specific for objects (e.g. units and events) and for variables (see table 3). Each dimension is briefly discussed below.

Technical checks: A total of 4 indicators are included in the Technical checks dimension. Apart from

Table 1: Quality framework for secondary data sources, Source hyperdimension

Dimensions	*Quality indicators*	*Methods*
1. Supplier	1.1 Contact	-Name of the data source
		-DSH[1] contact information
		-NSI[2] contact person
	1.2 Purpose	-Reason for use of the data source by DSH
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 & security	3.1 Legal provision	-Basis for existence of data source
	3.2 Confidentiality	-Does the Personal Data Protection Act apply?
		-Has use of 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 DSH
	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	-Contact DSH 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

[1] DSH: Data Source Holder; [2] NSI: National Statistical Institute.

indicators related to the accessibility and correct conversion of the data, this dimension also contains an indicator that checks if the specific data delivery complies to its metadata-definition. The metadata can be included in the delivery, either as a separate file or as a header in the file (describing its content), or as part of a separate delivery. In the Technical checks dimension also an indicator is included that expresses the results of preliminary data exploration. An example of this is shown in the paper by Tennekes et al. (2011).

Accuracy: Nearly all of the indicators in this dimension originate from the sources of error scheme when using administrative data (Bakker, 2010). The extended scheme of Zhang (2011) was used to derive indicators up to the point at which the data is linked to other (statistical) data sources. In the Accuracy dimension, the indicators for objects point to the correctness of the objects in the source, while the variable indicators focus on the validity of the values provided. Of the indicators, 4 are for objects and 5 for variables.

Completeness: The indicators for objects in this dimension predominantly focus on coverage issues. The indicators for variables in Completeness are related to missing and imputed values. Of the total of 6 indicators, 4 are object and 2 are variable specific.

Time-related dimension: The quality indicators in this dimension are all related to time. The timeliness

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

Dimensions	*Quality indicators*	*Methods*
1. Clarity	1.1 Population unit 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 unit definition comp.	-Comparability with NSI definition
	2.2 Classification variable def. comp.	-Comparability with NSI definition
	2.3 Count variable definition comp.	-Comparability with NSI definition
	2.4 Time differences	-Comparability with NSI reporting periods
3. Unique keys	3.1 Identification keys	-Presence of unique keys
		-Comparability with unique keys used by NSI
	3.2 Unique combinations	-Presence of useful combinations of variables
4. Data treatment (by DSH)	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

and punctuality indicators apply to the delivery of an individual data file. In addition, an indicator is included for the overall time lag. This indicator measures the laps of time between the end of the reference period covered by the data and the moment at which it can definitely be used by the NSI. As such, it also includes the time required for evaluation. The remaining 2 indicators are stability related. The object-indicator focuses on the usefulness of the source to (quickly) identify population changes while the variable-indicator looks at consistency of the values of specific variables over time. Although the values of variables will of course change between subsequent deliveries, it is important that the variable composition covered by a source remains stable and that the values of its variables (such as the NACE code of a company) show plausible changes between deliveries (Daas et al., 2011).

Integrability: This dimension contains indicators specific for the ease by which the data in the source can be integrated into the statistical production system of an NSI. The indicators for objects look at the comparability and link-ability of the objects in the source to those commonly used by the NSI. The variable indicators either focuses on the quality of the linking variable(s) used or compare the closeness of the values in the source to the values of the same or similar variables in other sources. A total of 4 indicators are included in this dimension, 2 for objects and 2 for variables.

3. Application

The framework described above is used to determine the *input* quality of administrative and other secondary data sources for statistics. The quality is determined by successively evaluating the quality components in the Source, Metadata, and Data hyperdimension (Daas et al., 2009; 2011). This strict order is the result of the fact that the quality assessed with the three hyperdimensions moves from a general to a more detailed level. This approach prevents that the user invests considerable time and effort in the study of quality components that are not relevant at that specific point in time. When the results for some of the quality indicators in a hyperdimension reveal problems, it is recommended to sort these out before the start of the evaluation of the next hyperdimension. This prevents that problems observed earlier on in the evaluation are (later on) found to be so sever that they prevent the use of the data source for the statistical

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

Dimensions	*Quality indicators*	*Methods*
1. Technical checks	1.1 Readability	-Accessibility of the file and data in the file
	1.2 File declaration	-Compliance of the data to metadata agreements
	1.3 Convertibility	-Conversion of the file to the NSI-standard format
	1.4 Data inspection	-Results of preliminary data analysis
2. Accuracy	*Objects*	
	2.1 Identifiability	-Correctness of identification keys for objects
	2.2 Authenticity	-Correspondence of objects
	2.3 Consistency	-Overall consistency of objects in source
	2.4 Dubious objects	-Presence of untrustworthy objects
	Variables	
	2.5 Validity	-Correctness of measurement method used by DSH
	2.6 Reporting error	-Errors made by the data provider during reporting
	2.7 Registration error	-Errors made during data registration by DSH
	2.8 Processing error	-Errors made during data maintenance by DSH
	2.9 Dubious values	-Presence of inconsistent combinations of values
3. Completeness	*Objects*	
	3.1 Undercoverage	-Absence of target object in the source
	3.2 Overcoverage	-Presence of non-target objects in the source
	3.3 Selectivity	-Statistical coverage and representativiness of objects
	3.4 Redundancy	-Presence of multiple registrations of objects
	Variables	
	3.5 Missing values	-Absence of values for (key) variables
	3.6 Imputed values	-Values resulting from imputation actions by DSH
4. Time-related dimension		
	4.1 Timeliness	-Time between end of reference period and receipt of source
	4.2 Punctuality	-Time lag between the actual and agreed delivery date
	4.3 Overall time lag	-Overall time difference between end of reference period and the moment NSI concluded that the source can be used
	Objects	
	4.4 Object dynamics	-Usefulness of source to identify population changes
	Variables	
	4.5 Variable stability	-Consistency of variables or values over time in source
5. Integrability	*Objects*	
	5.1 Object comparability	-Similarity of objects in source with the NSI-objects
	5.2 Alignment	-Linking-ability of objects in source with NSI-objects
	Variables	
	5.3 Linking variable	-Usefulness of linking variables (keys) in source
	5.4 Variable comparability	-Proximity (closeness) of variables

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 cannot be used for statistics at all. Because of their more application related focus, the indicators in the Metadata and Data hyperdimension need to be reviewed with a statistical use in mind (Daas et al., 2009; Schulte Nordholt et al., 2011).

The checklist (Daas et al., 2009) guides the user through the measurement methods for each of the quality indicators in the Source and Metadata hyperdimensions. By answering the questions in the checklist, the 'value' of every method for each indicator in tables 1 and 2 is determined. Since the predominant part of

the methods in the Source and Metadata hyperdimension are qualitative, usually a score has to be filled in. When problems are found or a question cannot be answered completely, the checklist guides the user in the steps to take. Additional space is included to write down remarks. Quite a number of data sources have been studied in this way. These include administrative sources on persons and enterprises (Daas et al., 2009; Schulte Nordholt et al., 2011) and some other secondary data sources, such as internet (Ossen et al., 2010) and secondary surveys (Daas et al., 2008a). These experiences demonstrate the general applicability of the framework and reveal that it takes, on average, around 2 hours to fill in the checklist. Current studies focus on the evaluation of the indicators in the Data hyperdimension (Daas et al., 2011). So far, only a limited number of indicators and data sources have been evaluated (Schulte Nordholt et al., 2011). The work on this topic is performed as part of the BLUE-Enterprise and Trade Statistics project in which the NSI's of the Netherlands, Norway, Sweden and Italy join forces (Daas et al., 2011).

Acknowledgment

Part of the work described in this paper was developed in the 7th Framework project BLUE-Enterprise and Trade Statistics. The authors thank Li-Chun Zhang, Coen Hendriks, Kristin Foldal Haugen (Statistics Norway), Antonio Bernardi, Fulvia Cerroni (Statistics Italy), Thomas Laitila, Anders Wallgren, and Britt Wallgren (Statistics Sweden) for their valuable contribution to the composition of the Data hyperdimension.

REFERENCES

Bakker, B. (2010) Micro-integration: State of the Art. Paper for the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, The Hague, The Netherlands.

Daas, P.J.H., Arends-Tóth, J., Schouten, B., Kuijvenhoven, L. (2008a) Proposal for a Quality Framework for the Evaluation of Administrative and Survey Data. Paper for the CENEX-workshop on the Combination of surveys and administrative data, Vienna, Austria.

Daas, P.J.H., Arends-Tóth, J., Schouten, B., Kuijvenhoven, L. (2008b) Quality Framework for the Evaluation of Administrative Data. Paper for the European Conference on Quality in Official Statistics 2008, Rome, Italy.

Daas, P.J.H., Ossen, S.J.L., Tennekes, M. (2010) The determination of administrative data quality: recent results and new developments. Paper for the European Conference on Quality in Official Statistics 2010, Helsinki, Finland.

Daas, P.J.H., Ossen, S.J.L., Tennekes, M., Zhang, L-C., Hendriks, C., Foldal Haugen, K., Bernardi, A., Cerroni, F., Laitila, T., Wallgren, A., Wallgren, B. (2011) List of quality groups and indicators identified for administrative data sources used as input in the statistical process. First deliverable of workpackage 4 of the BLUE-ETS project, March 10.

Daas, P.J.H., Ossen, S.J.L., Vis-Visschers, R.J.W.M., Arends-Toth, J. (2009) Checklist for the Quality evaluation of Administrative Data Sources. Discussion paper 09042, Statistics Netherlands.

Frost, JM, Green, S., Pereira, H., Rodrigues, S., Chumbau, A., Mendes, J. (2010) Development of quality indicators for business statistics involving administrative data. Paper for the European Conference on Quality in Official Statistics 2010, Helsinki, Finland.

Ossen, S.J.L., Daas, P.J.H., Puts, M. (2010) Quality framework for registers applied to online price information and offline route information. Paper for the European Conference on Quality in Official Statistics 2010, Helsinki, Finland.

Schulte Nordholt, E., Ossen, S.J.L., Daas, P.J.H. (2011) Research on the quality of registers to make data decisions in the Dutch Virtual Census. Paper for the 58th Session of the International Statistical Institute, Dublin, Ireland.

Tennekes, M., De Jonge, E., Daas, P.J.H. (2011) Visual profiling of Large Statistical Datasets. Paper for the 2011 European New Techniques and Technology for Statistics conference, Brussels, Belgium.

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

Wallgren, A., Wallgren, B. (2007) Register-based Statistics: Administrative Data for Statistical Purposes. Wiley Series in Survey Methodology, John Wiley & Sons, Ltd, Chichester, England.

Zhang, LC. (2011) Developing statistical theories for register-based statistics. Qvintensen 4, pp. 20 - 22, Swedish Statistical Association.

Overall Assessment of the Quality of Administrative Data Sources

3 high level views on input quality

	Exam Result Register	Higher Edu.	Soc. General Edu.	VAT
1. Supplier	–	+	+	–
2. Relevance	o	!	!	
3. Privacy and security		!	!/o	
4. Delivery	–	+	o	–
5. Processing	!/o	/o	!/o	o/-

	Exam Result Register	Higher Edu.	Soc. General Edu.	VAT
1. Liberty	!!	–	–	–
2. Comparability	–	–	–	-/o
3. Unique keys				o
4. Data treatment	/(o)	!/-(-)	/(+)	-/o

SOURCE
Focus on data source as a whole
- Contact information related
- Delivery related aspects
- and more

METADATA
Focus on the (availability of the) information required to understand and use the data in the data source

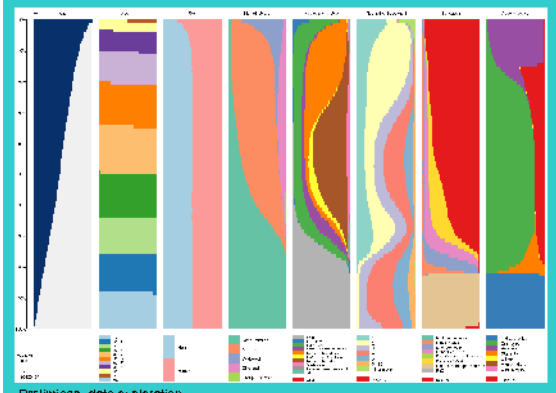
SOURCE

METADATA

DATA

DATA
Focus on the quality of the data

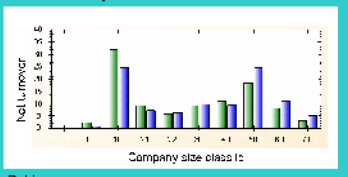
1. Technical checks



Preliminary data exploration

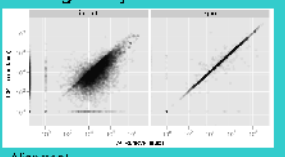
Available in
http://cran.r-project.org : packages : tabo ot

2. Accuracy



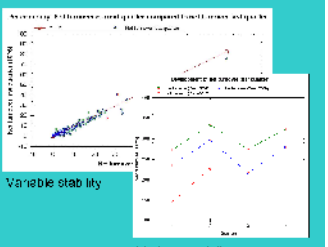
Dubious values

5. Integrability



Alignment

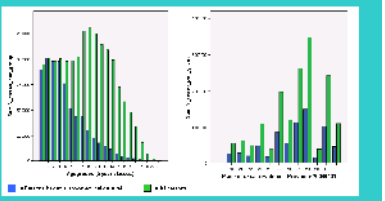
4. Time related



Variable stability

Variable stability

3. Completeness



Undercoverage/ selectivity

Saskia Ossen sjl.ossen@cbs.nl
Piet Daas pjh.daas@cbs.nl
Martijn Tennekes m.tennekes@cbs.nl
Statistics Netherlands, Division of Methodology and Quality
Centraal Bureau voor de Statistiek