Determination of Administrative Data Quality: Recent results and new developments

Piet J.H. Daas*, Saskia J.L. Ossen, and Martijn Tennekes
Statistics Netherlands, Division of Methodology and Quality
CBS-weg 11, 6412 EX
Heerlen, The Netherlands
pjh.daas@cbs.nl

Keywords: Register quality, Administrative data, Register-based statistics

1. Introduction

Many National Statistical Institutes (NSI's) want to increase the use of administrative data (such as registers) for statistical purposes. To enable the use of administrative data sources by NSI's a number of prerequisites have to be met. These are in decreasing order of importance:

1. Availability of administrative data sources
2. Conformation of the NSI to a set of preconditions to enable the use of administrative data sources on a regular basis
3. Availability of methods to evaluate the statistical usability (i.e. quality) of administrative data sources in a standardized way.

Point 1. To enable the use of administrative data for statistical purposes, relevant administrative data sources should of course be available in the home country of the NSI. Because of the increase in the use of information and communication technology in public administrations (e-Government), this should not be a problem in most countries nowadays (Socitm, 2002). One may expect that at least some of these data sources are of potential interest for NSI's (Wallgren and Wallgren, 2007).

Point 2. The second issue is the topic of an excellent 'best practice paper' by the NSI's of the Nordic countries (Unece, 2007). This paper gives a thorough overview of the preconditions required to enable an NSI to extensively make use of administrative sources in statistics production. Conformance to these conditions will enable an NSI to use administrative data for statistics on a regular basis.

Point 3. When the two prerequisites described above are met, the statistical usability of data sources becomes the dominant factor. It is essential that an NSI is able to determine the statistical usability (i.e. the quality) of the data source prior to use. Quality determination is an important issue because the collection and maintenance of an administrative data source is beyond the control of an NSI. It is the data source keeper that manages these aspects. The same holds for the units and the variables the data source contains. They are defined out of administrative rules and may therefore not be identical to those required by the NSI (Wallgren and Wallgren, 2007). For an NSI, it often takes considerable effort to unambiguously determine the usability of administrative data (ESC, 2007; Bakker et al., 2008).

1.1 Quality determination

The study of the quality of administrative and other secondary data sources would be greatly stimulated when a method would be available that is capable of determining the quality of these type of data sources -for statistical use- in a straightforward and standardized way. Another factor of importance is the time required for this; preferably evaluation times should be reduced to a

* The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. Part of the work described in this paper was performed under the BLUE-ETS project, a project that is financed by the 7th Framework Programme (FP7) of the European Union under the cooperation programme for Socio-economic Sciences and the Humanities, Provision for Underlying Statistics.

minimum. In the ideal situation, an NSI should be able to quickly get on overview of the quality of a data source. This not only requires reliable and efficient methods but also necessitates the availability of a 'well thought out' structured approach. As yet, however, no standard (and valid) instrument or procedure is available for this purpose. The development of such an instrument is the topic is of this paper.

2. Quality determination of administrative data sources

Quality is a multidimensional concept (Batini and Scannapieco, 2006). Ever since the adoption of the European Statistics Code of Practice, quality has become even more important for NSI's (Eurostat, 2005a). As a result of this adoption, NSI's of European Union (EU) member states have committed themselves to an encompassing approach towards high quality statistics. NSI's of the EU-member states and NSI's of some other European countries, such as Norway, report the quality of their statistical output by using six quality dimensions. These are: Relevance, Accuracy, Timeliness and punctuality, Accessibility and clarity, Comparability, and Coherence (Eurostat, 2005b). Both data and metadata related quality aspects are included in this framework (Batini and Scannapieco, 2006) that was specifically developed for the statistical output of NSI's (Eurostat, 2003a). For the determination of the quality of secondary data sources used for the input of NSI's the six quality dimensions can not be directly applied; see Eurostat (2003b) and Daas et al. (2008) for more details on this topic.

Much of the work of the authors of this paper has initiated from this starting point. In the beginning of our research, the various quality aspects of administrative data sources were identified (Daas et al., 2008) after which a structured way was developed to determine the metadata related quality aspects (Daas et al., 2009a). This paper starts with an overview of the former after which recent results of the latter are shown. At the end of the paper the next logical step is discussed and illustrated with examples

2.1 Quality framework: an overview

An extensive literature study revealed that the views on the composition of the quality of secondary data sources -to be used for statistics- varied greatly between papers (Daas et al., 2008). By combining the various quality aspects identified in each paper a complete overview of all the quality aspects of secondary data relevant for statistical use was obtained. During this work two important findings emerged. The first one is the fact that there is a clear 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 multidimensional concept. Depending on the perspective from which the data source is looked upon, different quality aspects prevail. Such a perspective -a high level view at statistical quality- is nothing new, it has been described several times by others. These views are usually called categories (Batini and Scannapieco, 2006) or hyperdimensions (Karr et al., 2006). The latter term is used in this paper.

A hyperdimension is a way of looking at quality at a level higher than that of the commonly used dimension (see above). For administrative data three hyperdimensions were identified, these are called Source, Metadata, and Data (Daas et al., 2008). The combined set of indicators in the Source and Metadata hyperdimension contains all quality indicators specific to the metadata domain of quality. The reader is referred to the papers of Daas et al. (2009a and 2009b) for a detailed overview of the indicators in the Source and the Metadata hyperdimension. The quality indicators for the data domain of quality are included in the Data hyperdimension (see table 3 below).

2.1.1 Source and Metadata hyperdimension: Exploratory phase

An NSI that plans to use a secondary data source should start by exploring the quality aspects of the information required to use the data source on a regular basis. These aspects of quality are located in the Source and Metadata hyperdimension of the quality framework (Daas et al., 2009a). In the Source hyperdimension, the predominant part of the quality aspects included relate to the undisturbed delivery of the data source -by the data source keeper- to the NSI. The quality aspects

in the Metadata hyperdimension focus on the conceptual and process related quality aspects of the metadata of the source. The conceptual metadata aspects check if the NSI is able to fully comprehend the definitions of the units, variables, and reporting period(s) of the data source keeper and compares them with those used by the NSI. The small number of process related indicators in the Metadata hyperdimension relate to the checks and modifications of the data by the data source keeper. It is important that the NSI is aware of any modifications of the data by the data source keeper because these highly affect the quality of the data in the source.

A checklist has been developed to assist the evaluation of the quality aspects in the Source and Metadata hyperdimension (Daas et al., 2009b). It is composed of questions that correspond to the measurement method(s) of every quality aspect in the Source and Metadata hyperdimension. The checklist also serves as a guide through the measurement methods. To test the usability of the framework and the checklist, eight secondary data sources of Statistics Netherlands were evaluated. The data sources studied were: Insurance Policy record Administration (IPA), Student Finance Register (SFR), register of the Centre for Work and Income (CWI), Exam Results Register (ERR), the coordinated register for Higher Education (1FigHE), the coordinated register for Secondary General Education (1FigSGE), the National Car Pass register (NCP), and the Dutch Municipal Base Administration (MBA). Due to page restrictions a detailed description of the data sources is not included in this paper.

In table 1 and 2 the dimensional evaluation results for the Source and Metadata hyperdimension are, respectively, shown. For the evaluation of the Metadata hyperdimension it is required that the user has a particular statistical use for the data source in mind (Daas et al., 2009b). This is due to the fact that the conceptual metadata definitions of units and variables in the data source need to be compared with those of the NSI (Daas et al., 2008). For the IPA, the Metadata hyperdimension part of the checklist was filled in with its use for the labour statistics in mind. The MBA was reviewed as a source for the population statistics and the NCP was evaluated with its use for the traffic and trade statistics in mind. The envisaged use of the other data sources was educational statistics (Bakker et al., 2008).

Evaluation scores are indicated at the dimension level. The symbols for the scores used in table 1 and 2 are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combing symbols with a slash (/) as separator. Since each dimension contains several quality indicators which are measured by one or more methods (Daas et al., 2009b), the results shown were obtained by comparing the evaluation results for every measurement method of each quality indicator in each dimension. Usually, the most commonly observed score was selected for the whole dimension. However, when in a dimension an unclear score occurs for a specific quality indicator, this score is shown for the whole dimension. Only when the scores for the other indicators in that dimension are not unclear, the most commonly observed score for these indicators is additionally shown between brackets.

The evaluation results for the eight data sources reveal that attention should be paid to the delivery, clarity, and comparability dimensions of the CWI before any studies are performed that relate to the data in this source. Only when the problems in the Source and Metadata hyperdimension of quality for the CWI are solved satisfactorily, it makes sense to spend (a lot more) time and effort in the determination of the quality of the data of this source. The results for MBA demonstrate that Statistics Netherlands currently has every quality aspect identified for the

Table 1. Evaluation results for the Source hyperdimension

Dimensions	Data Sources							
	IPA	SFR	CWI	ERR	1FigHE	1FigSGE	NCP	MBA
1. Supplier	+	+	+	+	+	+	+	+
2. Relevance	+	+	+	o	+	+	+	+
3. Privacy & security	+	+	+	+	+	+/o	+	+
4. Delivery	o	+	-	+	+	o	+	+
5. Procedures	+	+/o	+	+/o	+/o	+/o	o	+

Table 2. Evaluation results for the Metadata hyperdimension

Dimensions	Data Sources							
	IPA	SFR	CWI	ERR	1FigHE	1FigSGE	NCP	MBA
1. Clarity	+	+	-	o	+	+	+	+
2. Comparability	+/o	+	-	+	+	+	+	+
3. Unique keys	+	+	+	+	+	+	+	+
4. Data treatment	+/o	?(+)	?	?(o)	?(+)	?(+)	+	+

Source and Metadata hyperdimension under control. For the other data sources it can be argued that the results suggest that some of the quality aspects in some or both hyperdimensions require attention. But overall no serious problems were found.

The results described above demonstrate that the checklist, and therefore also the quality aspects in the Source and Metadata hyperdimensions, are valuable for determining the quality of the prerequisites for use of a secondary data source for statistics. For all data sources listed above, with exception of the CWI, the next step would be the study of the quality aspects of the Data hyperdimension (Daas et al., 2008 and 2009b).

2.2 Data quality of administrative data sources

When for an administrative data source no substantial problems are observed during the exploratory phase of quality (see above) an NSI should start looking at the quality of the data in the source (Daas et al, 2008). The quality indicators identified for the Data hyperdimension are listed in table 3. These are also the result of the literature study mentioned above (see Daas et al., 2008 for more details). The quality aspects included in table 3 relate to the dataset as a whole, the units, and the facts in the data source.

Many of the quality indicators in table 3 are familiar to statisticians, some are probably not. The latter will be briefly discussed here. A considerable part of the measurement methods listed in table 3 is based on the so-called Representativity index (R-index). This indicator was developed at Statistics Netherlands (Schouten et al., 2009). 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 (Schouten et al., 2009). 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 Value-added tax (VAT) data to the Dutch Tax Office varies during the monthly collection period (Ouwehand et al., 2009). This affects the quality of the data provided to our office. 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. In table 3 also the dimension Precision is included. For administrative data, this dimension is mainly used to determine the effect of time-dependent changes in the population composition on data quality.

2.2.1 Structured study of data quality

When studying the quality of the data of administrative data sources quite some quality aspects will have to be determined. The list in table 3 already contains 33 measurement methods. It is presumably not very efficient for an NSI to measure the value for each indicator every time a new data source is received. Let alone the amount of work for data sources that arrive bit by bit, such VAT-data. In all cases, only the essential and strictly required checks and indicators should be determined first. When problems are observed at that stage, more detailed studies are required. This pragmatic 'point-of-departure' is the basis for the quality determination process shown in figure 1.

For completeness, this figure also contains the checklist for the quality aspect included in the Source and Metadata hyperdimension used in the exploratory phase. The subdivision of the quality aspects in the Data hyperdimension is the most important part of figure 1. In the study of the quality of the data, three stages can be discerned. These stages differ greatly in their level of effort and

Table 3. Dimensions, quality indicators, and description of methods proposed for Data

DIMENSIONS	QUALITY INDICATORS	METHOD DESCRIPTION
1. Technical checks	1.1 Readability	-Can all the data in the source be accessed?
	1.2 Metadata compliance	-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
3. Under coverage	3.1 Missing units	-Percentage of units missing from the target population
	3.2 Selectivity	-R-index a) for unit composition
	3.3 Effect on average	-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 unambiguously
	4.2 Mismatches	-Percentage of units incorrectly linked
	4.3 Selectivity	-R-index for composition of units linked
	4.4 Effect on average	-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 with all data missing
	5.2 Selectivity	-R-index for unit composition
	5.3 Effect on average	-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 average	-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 bias (relative measurement error)
8. Processing	8.1 Adjustments	-Fraction of fields adjusted (edited)
	8.2 Imputation	-Fraction of fields imputed
	8.3 Outliers	-Fraction of fields corrected for outliers
9. Precision	9.1 Standard error	-Mean square error 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 totals
		-Maximum RMSE of totals

a) R-index: Representative Index, an indicator that estimates the selectivity of the data missing by using information available in other sources (Schouten et al., 2009). b) RMSE: root mean square error; a common used statistical measure for the quality of an estimator. The RMSE is equal to the square root of the sum of the bias and variance of the estimator.

detail. The stages are named: i) Technical checks, ii) Accuracy related quality indicators, and iii) Output related quality indicators. Each stage is discussed below.

2.2.2 Technical checks

In the first level, the technical checks-level, a quick check is performed of the data delivered to the NSI. The checks are very basic. An example of such a check is the comparison of the files size or the number of (unique) units in the data set to the size or number that is normally expected to be delivered. Another example of a technical-check is that of the compliance of the data to the metadata. This is a check that can be easily performed when the data is received in XML-format (Van der Vlist, 2002). At this stage it is important that the checks are able to quickly identify any serious errors.

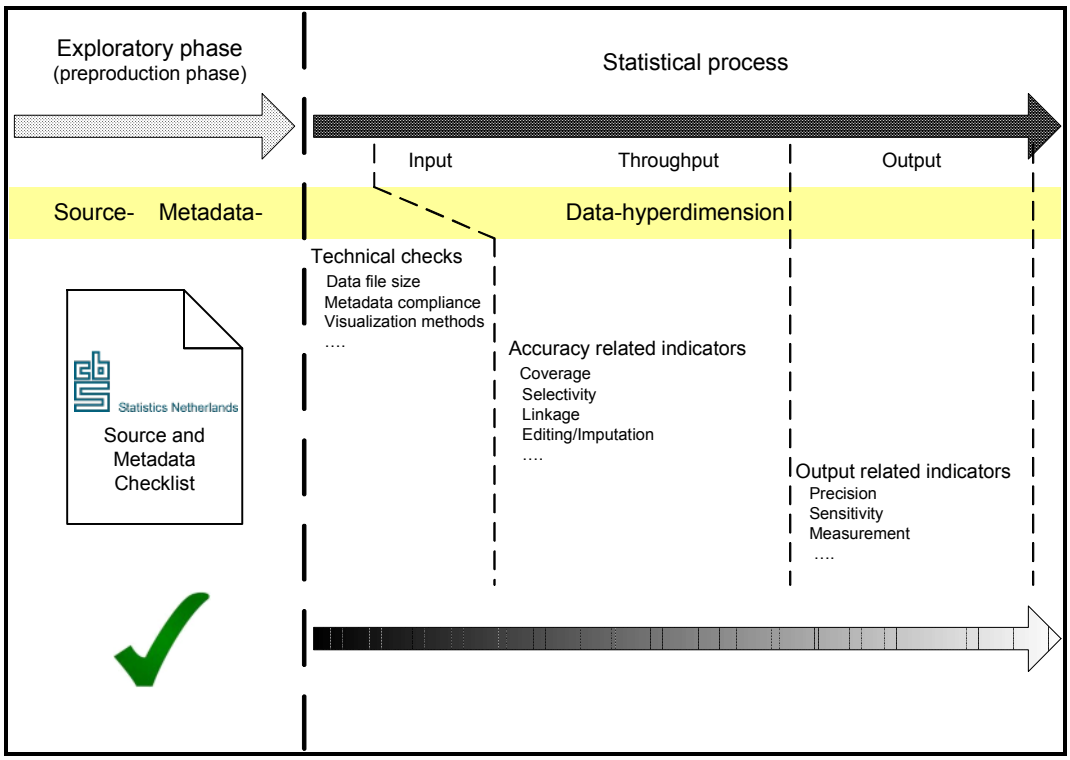

Figure 1. Overview of the process of quality aspect measurement for administrative data sources

Apart from the very basic examples given above, we think that more advanced checks can and should be used at this stage. Because computers can easily deal with large files and large numbers of records nowadays, the use of visualization methods is becoming a serious possibility at this stage in the process. Visualization methods can potentially greatly increase the added value of checks performed early on in the processing chain. It is already an approach that is used considerably in the data exploratory phase of large data sources (Pyle, 1999). Two examples of the use of visualization based checks are shown in figure 2 and 3 respectively. The first example is taken from Templ and Filzmoser (2008) and the second example is from the book of Uwin et al. (2006).

In figure 2 the result of a visualization method for missing data is shown. Here, the relation between the absence of the data for a number of variables is displayed in a graphical way. There are also other possibilities. A visualization method for missing data could also be used to observe patterns in the absent data. Imagine two data sources in which the same total percentage of facts are missing but for a different number of variables. These sources will display very different 'missing data' patterns when, for example, in the first source 10.000 units are (more or less randomly) missing the value for a single variable and in the second source 100 units are missing all the data for exactly 100 variables. The latter is (very likely) a reason to contact the data source keeper, while the former is probably not!

The example in figure 3 is from Uwin et al. (2006). It shows a table plot of aggregates of the first 12 variables in US Census data. Each column represents a variable and each row is an aggregate of 250 cases. The data has been sorted by age. A table plot is used to quickly get an overview of the (raw) data of a dataset. It tries to capture the information of the whole dataset in a single figure. What becomes apparent from figure 3 is that the proportion of males to females is higher for both older and younger groups in the dataset. Although it is likely that these types of plots will have to be tailor-made to each data source, the general principle could presumably be applied to (secondary) data sources in general. The use of visualization methods to obtain an overview of the data upon delivery is something that has to be seriously investigated for several (large) data sources in several countries. If visualization methods can be applied on a general basis this would enable the development of standardized methods and probably also of software tools to provide a quick 'peek' into the data. Such methods might very likely become part of a standardized

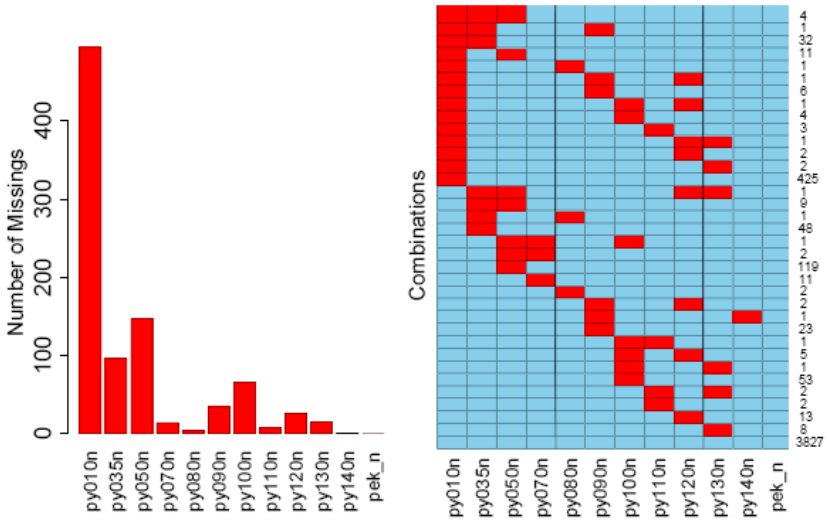

Figure 2. Example of the visualisation of missing data (from Templ and Filzmoser, 2008)

In the left part of the figure a bar chart is shown in which the number of missing values in a data source for 12 variables are plotted. In the right part of the figure the presence and absence of values for various combinations of those variables is shown. Each row represents a specific combination of variables with red and blue cells indicating the absence and presence of the value of the corresponding variable, respectively. The number on the right is the number of occurrences of this combination.

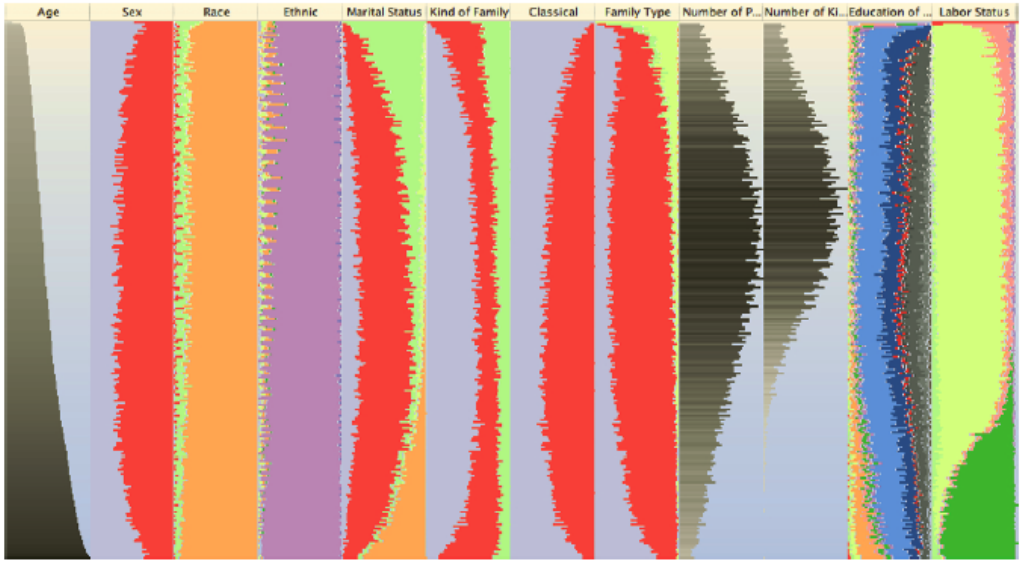

Figure 3. A table plot for aggregates of variables of US Census data (from Uwin et al., 2006)

In this figure the aggregated results of the first 12 variables of the data source are shown. Each column represents a variable, and each row is an aggregate of 250 units. The results are sorted according to age; the first column. Colours are used to differentiate between the various categories of categorical variables.

way to check a new data delivery to an NSI. Such a way of checking the data would not only be beneficial to NSI's but to all users of (large) data sources.

2.2.3 Accuracy related quality indicators

When an NSI has decided for which publication a data source will be used, more specific quality indicators can be applied. The indicators in this stage are called 'accuracy related indicators' because these types of indicators all, directly or indirectly, relate to the accuracy of the data. Many of the indicators listed in table 3 belong to this group. Examples of indicators for units in this group are: over- and under-coverage, selectivity (Ouwehand et al., 2009), and linkability. Examples of

indicators for the values of variables are: selectivity, the percentage of adjusted and imputed values, and external validation.

Many of the indicators in table 3 are already applied to survey data at Statistics Netherlands, which provides a solid basis for their potential use for administrative data. In addition, the general applicability of accuracy related quality indicators to both survey and administrative data is a very interesting added value. To what extent this is achievable in practice, however, remains to be seen (see below).

2.2.4 Output related quality indicators

The quality indicators in the third stage report on the quality of administrative data sources at an aggregated level. These indicators are all output oriented. They report on a level that, for an NSI, ultimately determines the statistical usability of a data source. The indicators in this group all try to answer the question: how good are the statistics produced from the data source? Examples of indicators belonging to this group are indicators that aim to determine: the precision of core variables and the selectivity of composite totals. Determination of the former indicator has made considerable progress during the last years when the combination of survey and administrative data sources is studied (Harmsen et al., 2009). However, determining them for aggregates fully based on administrative data is still a great challenge (Zhang, 2009). The study of the selectivity of units also has made great progress since the start of the Representative Indicators for Survey Quality (RISQ) project (Schouten et al., 2008). Although the selectivity indicators studied in this project were originally developed for survey data, it was recently demonstrated that these indicators can also be applied to administrative data sources (Ouwehand et al., 2009).

There is however a restriction to the indicators in this stage, and probably also to some of those included in the previous stage. It is important to realize that the quality indicators included in the framework should and need to be generally applicable. Very specific indicators can not be included, simply because it is impossible to include all possible conceivable indicators (Daas et al., 2008). The following illustrates this very important point. An example of a very specific indicator 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. It is clear that such an indicator can and should not be included in the set of indicators for data because it can not 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 scope of the framework has to be restricted to some extent as it is impossible to meet all conceivable uses.

Attention also needs to be paid here to the question whether the indicators in this group provide information on the usability of the data source prior to use. When an indicator can only be measured after the data in the source has been processed, the indicators can not be used as an early indicator for the quality of the source. For that purpose the checks and indicators in the first two stages are obviously better suited.

2.2.5 Implementation of data quality determination

Studies of the quality indicators in the Source and Metadata hyperdimension have shown that implementation of these indicators is greatly enhanced when it is embedded in a structured approach. For the indicators in the Source and Metadata hyperdimension this was done by developing a checklist (Daas et al., 2009b). Application of the checklist revealed that -in the Netherlands- on average around 2 hours were needed to complete it (Daas et al., 2009a). The checklist, however, predominantly consists of qualitative questions, so usually a score had to be filled in.

For the indicators in the Data hyperdimension a checklist does not seem the most appropriate way. For instance, some of the checks in the technical checks part will very likely be based on a visual inspection. Add to this the fact that quite a lot of indicators and checks need to be evaluated for a data source and is becomes apparent that a checklist is very likely not the most efficient way to

determine data quality. Availability of standardized script and/or a software implementation of (a large part of) the checks and indicators in the Data hyperdimension are very likely more appropriate ways to assist the user. Robustness and ease of use (and dummy-proofness) of both will also have to be taken into account of course.

It is very important that the scores obtained for the quality indicators and checks in the Data hyperdimension are combined into a single instrument. This to assure the central availability of the quality results for an administrative data source. Such an instrument, e.g. a Quality Report Card or Quality Report Tool, should provide an overview of all the quality scores essential for the data source under study and should be generally applicable to all administrative data sources in as many countries as possible. Another major challenge is the normation of the scores obtained. Compared to the (almost entirely) qualitative results in the Source and Metadata hyperdimension (Daas et al., 2009b), the establishment of norms is much more challenging for the predominantly quantitative results in the Data hyperdimension (see table 3).

3. Conclusions

In this paper an overview is given of the current state of art and ideas for the study of the quality of administrative data sources by NSI's. The basis of the work described has been laid down by Statistics Netherlands but more effort is required to finalize it. The recently started BLUE Enterprise and Trade Statistics (BLUE-ETS) project, a project financed by the 7th Framework Programme (FP7) of the European Union under the cooperation programme for Socio-economic Sciences and the Humanities, brings together statisticians from 14 institutes to, amongst others, pick up some of the tremendous key challenges in official business statistics. One of the goals of this project is to stimulate the use of administrative data sources for statistics by developing a procedure to determine the usability of administrative data sources in an efficient and straightforward way. In achieving this goal, the authors of this paper will be involved together with statisticians from the National Statistical Institutes of Italy, Norway, Slovakia, and Sweden. This research will eventually result in the development of a new comprehensive quality-indicator instrument for administrative registers. This paper describes the starting point and points the road ahead for the research required to achieve this goal.

REFERENCES

Bakker, B.F.M., Linder, F., Van Roon, D. 2008. Could that be true? Methodological issues when deriving educational attainment from different administrative datasources and surveys. Proceedings of IAOS Conference on Reshaping Official Statistics, Shanghai, China.

Batini, C., Scannapieco, M. 2006. Data Quality: Concepts, Methodologies and Techniques. Springer, Berlin

Daas, P.J.H., Arends-Tóth, J., Schouten, B., Kuijvenhoven, L. 2008. Quality Framework for the Evaluation of Administrative Data. In: Proceedings of Q2008 European Conference on Quality in Official Statistics. Statistics Italy and Eurostat, Rome.

Daas, P.J.H., Ossen, S.J.L., Arends-Tóth, J. 2009a. Framework of Quality Assurance for Administrative Data Sources. Paper for the 57th session of the International Statistical Institute, 16-22 Aug., Durban, South Africa.

Daas, P.J.H., Ossen, S.J.L., Vis-Visschers, R.J.W.M., Arends-Toth, J. 2009b. Checklist for the Quality evaluation of Administrative Data Sources. Discussion paper 09042, Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/0DBC2574-CDAE-4A6D-A68A-88458CF05FB2/0/200942x10pub.pdf

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, Geneva, Switzerland.

Eurostat 2002. Quality Declaration of the European Statistical System, Luxembourg.

Eurostat 2003a. Definition of quality in statistics, Assessment of the quality in statistics, Item 4.2, Luxembourg.

Eurostat 2003b. 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.

Harmsen, C., Van Der Laan, J., Kuijvenhoven, L. 2009. Deriving longitudinal consistent household statistics from register information. Paper for the 57th session of the International Statistical Institute, 16-22 Aug., Durban, South Africa.

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

Ouwehand, P., Schouten, B., De Heij, V. 2009. Representativity indicators for business surveys based on population totals. Paper for the European Establishment Statistics Workshop, 7-9 Sept., Stockholm, Sweden

Pyle, D. 1999. Data preparation for data mining. Morgan Kaufmann, San Francisco, USA.

Schouten, B., Cobben, F., Bethlehem, J. 2009. Indicators for the representativeness of survey response. Survey Methodology, 35 (1), 101-113.

Socitm 2002. Local e-government now: a worldwide view. Joint report of the Society of Information Technology Management and the Improvement & Development Agency, September.

Templ, M., Filzmoser, P. 2008. Visualization of missing values using the R-package VIM. Forschungsbericht CS-2008-1, Institut f. Statistik u. Wahrscheinlichkeitstheorie, Wien, Austria.

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

Unwin, A., Theus, M., Hofmann, H. 2006. Graphics of Large Datasets: Visualizing a Million. Spinger, Singapore.

Van der Vlist, E. 2002. XML Schema. O'Reilly & Associates Inc., Sebastopal, USA.

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, L-C. 2009. Unit errors in statistical registers and their effects. Paper for the 57th session of the International Statistical Institute, 16-22 Aug., Durban, South Africa.