Evaluation and visualisation of the quality of administrative sources used for statistics

Piet J.H. DAAS[*], Saskia J.L. OSSEN, Martijn TENNEKES and Joep M.S. BURGER

Statistics Netherlands, Department of Methodology and Development, P.O. Box 4481, 6401 CZ, Heerlen, The Netherlands; pjh.daas@cbs.nl

More and more National Statistical Institutes (NSI's) use administrative data as a primary source of information for producing statistics. Because the collection and maintenance of administrative data is beyond the control of NSI's, it is essential that NSI's are able to determine the quality (i.e. the statistical usability) of these sources when they enter the office in a quick, straightforward, and standardised way. The quality of the metadata components of administrative sources can be easily determined with the checklist developed by Statistics Netherlands. However, for the determination of the quality of the data in administrative sources a standard procedure was not yet available. This was the focus of research performed in part of the seventh framework project BLUE-Enterprise and Trade Statistics (BLUE-ETS). To enable the structured evaluation of administrative data, first the quality dimensions of administrative input data were identified: Technical checks, Accuracy, Completeness, Time-related, and Integrability. Next, for each dimension, quality indicators and measurement methods were developed which form the basis of the quality-indicator instrument to be tested in the remainder of the project. The overall approach is discussed and illustrated with examples.

1. Introduction

Many National Statistical Institutes (NSI's) want to increase the use of administrative sources (i.e. registers) for statistical purposes. This requires that relevant administrative sources need to be available in the home country of the NSI and that several preconditions have to be met [15]. The preconditions that enable an NSI to extensively make use of administrative sources in statistics production are: 1) legal foundation for the use of

[*] 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-Enterprise and Trade Statistics project (Grant Agreement Number 244767), a project 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.

administrative sources, 2) public understanding and approval of the benefits of using administrative sources for statistical purposes, 3) the availability of an unified identification system across the different sources used, 4) comprehensive and reliable systems in public administrations and 5) cooperation among the administrative authorities.

When the prerequisites described above are met, the statistical usability of administrative sources becomes an important issue. To cope with fluctuations in the quality of these sources, it is essential that an NSI is able to determine its statistical usability (i.e. the quality) on a regular basis. This is an important issue because the collection and maintenance of an administrative source are beyond the control of an NSI. It is the data source holder that manages these aspects. It is therefore of vital importance that an NSI has a procedure available that can be used to determine the quality of administrative sources for statistical use -when it enters the office- in a quick, straightforward and standardised way.

For the evaluation of the *metadata* quality components of administrative data sources a procedure [7] has been developed. This approach has been thoroughly evaluated in the Netherlands and proved very useful [5]. However, no standard instrument or procedure is available for the evaluation of the quality of the *data* in administrative sources when it *enters* the NSI [6]. This is commonly referred to as the input quality of administrative data [6,8]. The development of an approach to routinely evaluate the input quality of administrative data -for statistical purposes- is the main focus of Workpackage 4 (WP4) of the BLUE-Enterprise and Trade Statistics (BLUE-ETS) project [4]. The results of this workpackage form the basis of the work described in this paper.

2. Input quality of administrative data

2.1 Input quality components

By carefully reviewing the literature and discussing the findings with the other statisticians involved in the BLUE-ETS project, the key quality constituents of administrative *data* were identified. The work started by carefully studying how researchers in statistics and other research areas perceive and determine the quality of the secondary data sources they use as input for their work. This enabled us to identify the components of quality that are *generally* considered the most important by the users of secondary data [8]. The additional discussions in the group, including a comparison to current practices, resulted in the identification of five essential dimensions, namely: 1) Technical Checks, 2) Accuracy, 3) Completeness, 4) Time-related and 5) Integrability. For each dimension, relevant indicators were subsequently identified (Table 1).

For more details on the findings of the literature study performed and the process used to identify key dimensions and indicators the reader is referred to [8] and in particular to Annex A of [8]. Next, for every indicator, measurement methods have been developed [9]. At the time of writing, these methods are being implemented in R. For updates on the progress of this work, the reader is referred to the BLUE-ETS website [4].

2.2 On the determination of input quality

Determining the input quality of administrative data can be looked upon from two points of view [8,9]. The first one is the *data archive* point of view. From this general viewpoint the potential use of the data at the NSI may be anticipated, but it is only to a limited extent subject-specific. The other view on input quality is taken by a statistical user of the data who already has a *specific* use of the data in mind. The deficiencies and strengths of the data are weighted accordingly: certain deficiencies of the data may not be important, while others are critical. What is important to notice is that, even though the data is the same and the indicators are also all nearly the same (see below), the *assessment* of its quality may differ depending on the point of view taken. For example, suppose that an NSI obtains an administrative source containing only data for a selective part of the population (e.g. the small and medium sized enterprises) for which the data are technically correct, accurate, timely and can be integrated well. If the NSI plans to use this source on its own to produce Structural Business Statistics, then the data is obviously not good enough. However, if the source can be combined with, say, survey data collected from all large enterprises, then the administrative data source would be highly useful.

The relevancy of some of the indicators is also affected by the point of view taken [9]. The Integrability-related indicators (5.1 through 5.4) are clearly relevant for the specific -goal-oriented- view and are considered less important for the data archive point of view. The same reasoning applies to the stability-related indicators (4.5 and 4.6) of the Time-related dimension. In the Completeness dimension, the overcoverage indicator (3.2) is considered somewhat less relevant for the data archive point of view

Table 1. Quality dimensions and indicators for administrative input data used for statistics

Dimension Indicators	Description
1. Technical checks	*Technical usability of the file and data in the file*
1.1 Readability	Accessibility of the file and data in the file
1.2 File declaration	Compliance of the data in the file to the metadata
1.3 Convertability	Conversion of the file to the NSI-standard format
2. Accuracy	*The extent to which data are correct, reliable and certified*
Objects	
2.1 Authenticity	Legitimacy of objects
2.2 Inconsistent objects	Extent of erroneous objects in source
2.3 Dubious objects	Presence of untrustworthy objects
Variables	
2.4 Measurement error	Deviation of actual value from ideal error-free value, occurring during reporting, registration, or processing of data
2.5 Inconsistent values	Extent of inconsistent values for combinations of variables
2.6 Dubious values	Presence of implausible values or combinations of values
3. Completeness	*Degree to which a data source includes data describing the corresponding set of real-world objects and variables*
Objects	
3.1 Undercoverage	Absence of target objects (missing objects) in the source
3.2 Overcoverage	Presence of non-target objects in the source
3.3 Selectivity	Statistical coverage and representativity 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	Presence of values resulting from imputation actions by DSH[a]
4. Time-related dimension	*Indicators that are time and/or stability related*
4.1 Timeliness	Time lag between the end of the reference period in the source and the moment of receipt
4.2 Punctuality	Time lag between the settled date and actual delivery date
4.3 Overall time lag	Time lag between the end of the reference period in the source and the moment NSI concluded the data can be used
4.4 Delay	Time lag between an actual change in the real-world and its registration in the source
Objects	
4.5 Dynamics	Changes in the population of objects (births/deaths) over time
Variables	
4.6 Stability	Changes of variables or values over time
5. Integrability	*Extent to which the data source is capable of undergoing integration or of being integrated in the statistical system*
Objects	
5.1 Comparability of objects	Similarity of objects in source -at the proper level of detail- with objects used by NSI
5.2 Alignment of objects	Linking-ability (align-ability) of objects with those of NSI
Variables	
5.3 Linking variable	Usefulness of linking variables (keys) in source
5.4 Comparability of variables	Proximity (closeness) of variable values in different sources

[a] DSH, Data Source Holder

and the redundancy indicator (3.4) is considered somewhat less relevant for the goal-oriented point of view. For all other indicators and dimensions there are no differences.

3. Examples

For each dimension of administrative input quality typical examples will be discussed.

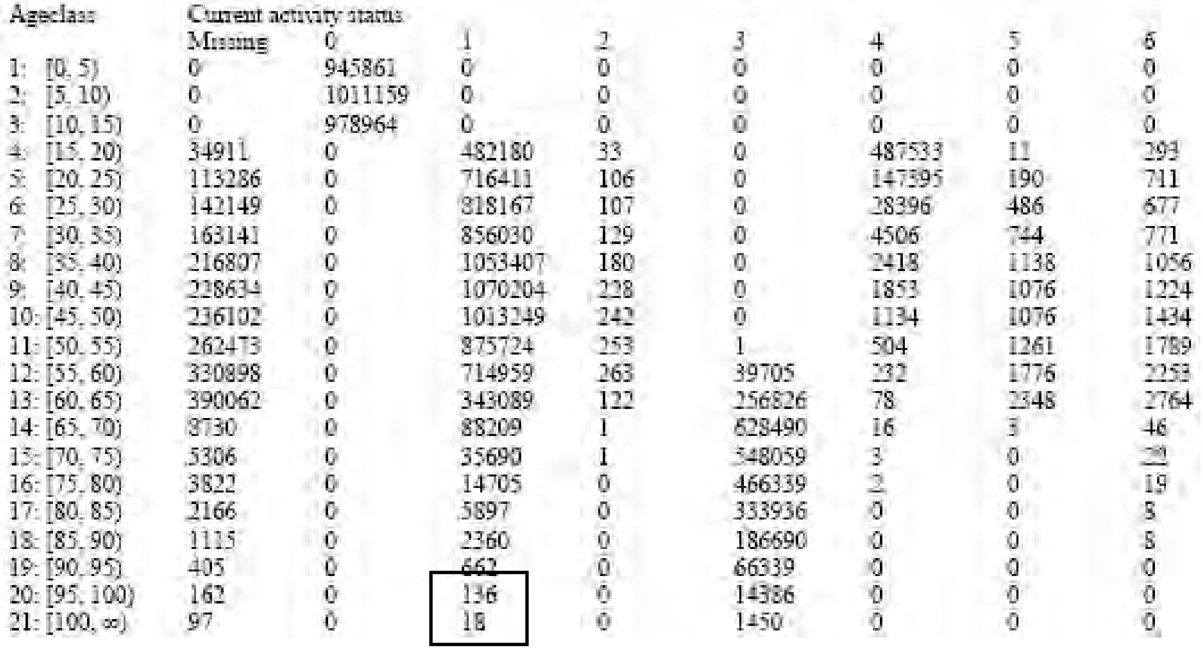
3.1 Technical checks

The Technical checks dimension predominantly consists of IT-related indicators for the data in a source. Apart from 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 with 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), but could also be provided to the NSI in a separate process. This approach very much resembles analysis commonly referred to as data profiling [10]. All indicators in this dimension focus on supporting the decision to either carry on using the data source or to report problems back to the data source holder. Such checks are especially important when a new data source is being studied, but become less important once routine use has come into place [9]. The end result of the Technical Checks dimension is essentially a go/no go decision.

3.2 Accuracy

The indicators in the Accuracy dimension all originate from the sources of error scheme published for administrative data [17]. This scheme identifies the sources of error when administrative data is used as input by NSI's up to the point at which the data is linked to other (statistical) data sources. One of the measurement methods of the authenticity indicator (2.1) checks the syntactical correctness of the identification number used [9]. In the Netherlands it was found, at the start of the Dutch Social Statistical Database, that 0.3% of all persons in the various administrative data sources used had an invalid Personal Identification Number [1]. Measurement errors (2.4) can be identified by comparing data registered for the same variables in several sources. A way to determine this is described by Bakker [3]. The dubious values indicator (2.6) is used to indicate the occurrence of implausible values or combinations of values. An example of this are the more then 150 employed people of age 95 and higher in the Dutch Virtual Census test file (Table 2).

Table 2. Cross tabulation of the variable ‘Current activity status’ by ‘age group’ in the Dutch Virtual Census test file [c]



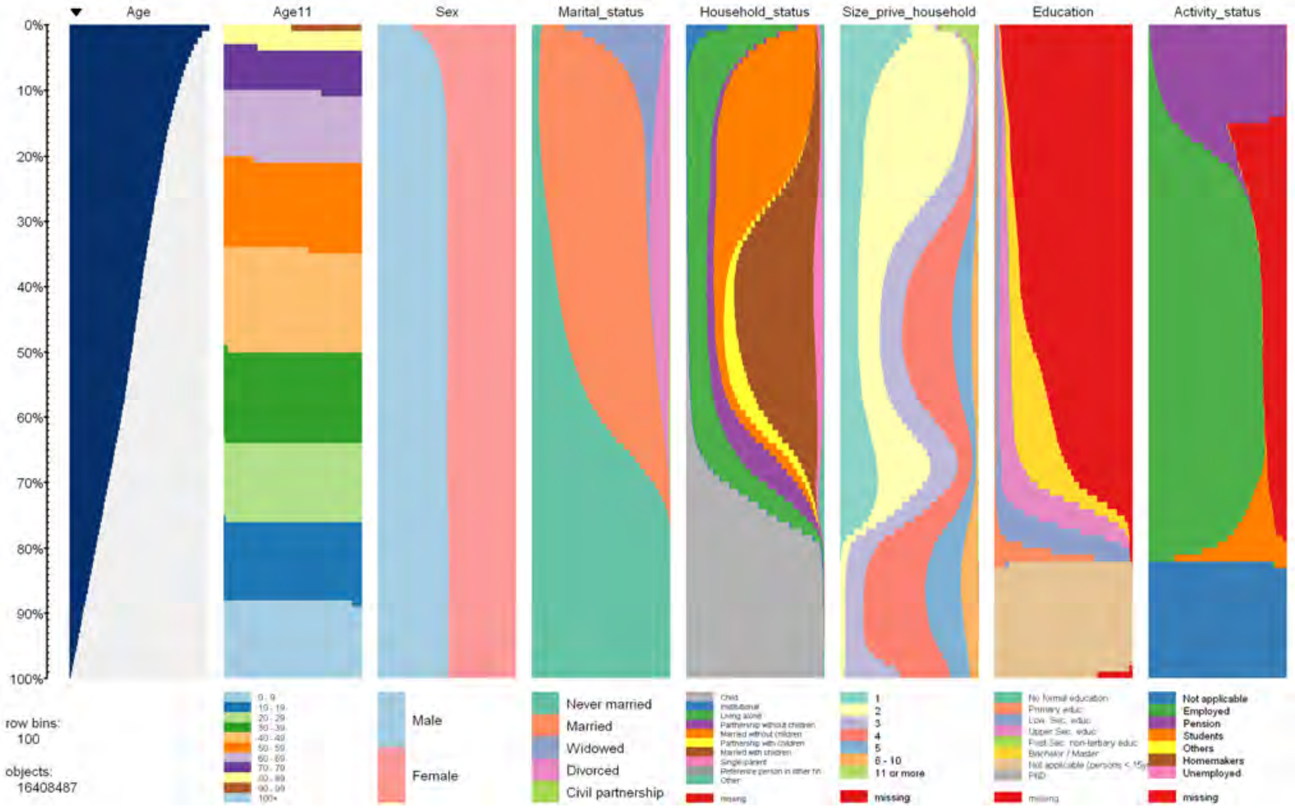
Ageclass	Current activity status							
	Missing	0	1	2	3	4	5	6
1: [0, 5)	0	945861	0	0	0	0	0	0
2: [5, 10)	0	1011159	0	0	0	0	0	0
3: [10, 15)	0	978964	0	0	0	0	0	0
4: [15, 20)	34911	0	482180	33	0	487533	11	293
5: [20, 25)	113286	0	716411	106	0	147595	190	711
6: [25, 30)	142149	0	818167	107	0	28396	486	677
7: [30, 35)	163141	0	856030	129	0	4506	744	771
8: [35, 40)	216807	0	1053407	180	0	2418	1138	1056
9: [40, 45)	228634	0	1070204	228	0	1853	1076	1224
10: [45, 50)	236102	0	1013249	242	0	1134	1076	1434
11: [50, 55)	262473	0	875724	253	1	504	1261	1789
12: [55, 60)	330898	0	714959	263	39705	232	1776	2253
13: [60, 65)	390062	0	343089	122	256826	78	2348	2764
14: [65, 70)	8730	0	88209	1	628490	16	3	46
15: [70, 75)	5306	0	35690	1	348059	3	0	22
16: [75, 80)	3822	0	14705	0	466339	2	0	19
17: [80, 85)	2166	0	5897	0	333936	0	0	8
18: [85, 90)	1115	0	2360	0	186690	0	0	8
19: [90, 95)	405	0	662	0	66339	0	0	0
20: [95, 100)	162	0	136	0	14386	0	0	0
21: [100, ∞)	97	0	18	0	1450	0	0	0

[c] Current activity status: (0) Persons below minimum age for economic activity, (1) Employed, (2) Unemployed, (3) Pension or capital income recipients, (4) Students not economically active, (5) Homemakers, (6) Others. The square indicates dubious values for employed people of age 95 and higher [12].

3.3 Completeness

The Completeness dimensions focuses on indicators for objects, which predominantly relate to coverage issues, and indicators for the values of variables, which relate to missing and imputed values. The selectivity indicator (3.3), for example, looks at the statistical coverage and representativity of objects (units) in the data source. An example of this is the age-related undercoverage found for the variable ‘level of education’ in the Dutch Virtual Census test file [12]. This can be calculated by the so-called R-indicator [11] but can also be illustrated by a visualization method specifically developed for the inspection of large data files; the so-called tableplot [13] which is shown in Figure 1. This tableplot visualizes the Virtual Census test file, displaying 8 variables for a total of 16.5 million records (all registered Dutch inhabitants in 2008) sorted by age. Each column represents a variable and each row (‘bar’) is an *aggregate* of a fixed number of records (here a percentile). The numeric sorting variable ‘age’ is displayed as a bar chart (in blue) and the other variables are categorized and shown as

Figure 1. Tableplot of the 2008 Dutch Virtual Census test file



stacked bar charts with a different colour for each category. The seventh column in Figure 1 displays the various categories for 'level of education' and illustrates the occurrence and distribution of missing values; shown in red. In the Netherlands, people over 15 can have various levels of education. However, with increasing age the amount of missing information increases dramatically. This is caused by the fact that the official registration of the level of education of graduates has only recently started in the Netherlands [2,12]. As a result, only the level of education of people that have recently finished school is stored in various public administrations which are predominantly young people. For all others, sample surveys are the only source of information in which this kind of information is available; explaining the increasing number of missing values with increasing age. Under 15, people do not have a formal level of education and should be categorized as 'not applicable'. The lowest two rows in the seventh column, however, clearly contain a considerable number of missing values. This is an obvious error that should be corrected. The eighth column also reveals that data on 'activity status' are also selectively missing for people between 15 and 65 years [12].

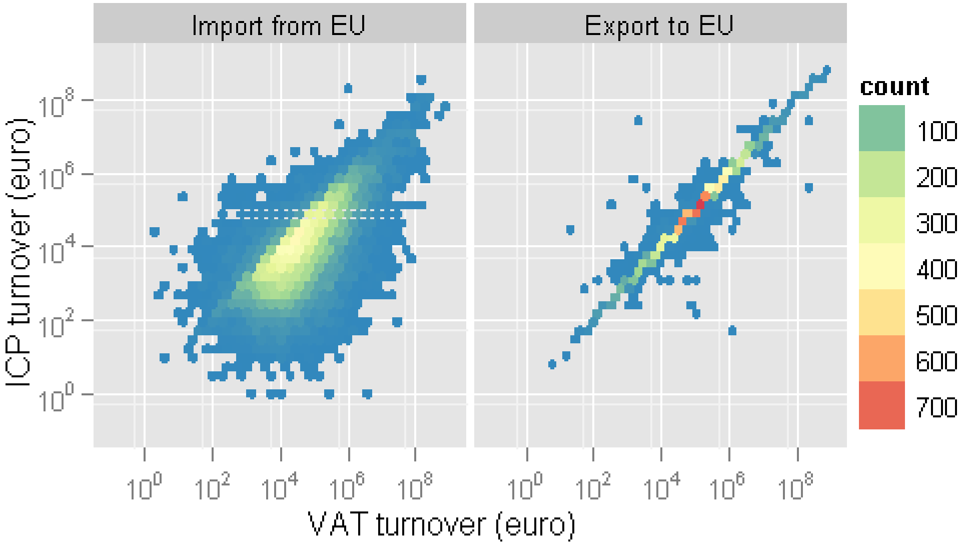
3.4 Time-related dimension

The indicators in the Time-related dimension focus on the time-related quality issues of individual data files. The delay indicator (4.4), for example, is used to indicate how quickly events are registered in the data source. Examples of problems in this area are: i) marriages contracted in immigrants' country of origin, which are sometimes recorded two or three years after the event [2]; ii) corrections in the Norwegian Population Register that continue to be reported over a lengthy period of time [16] and iii) Value Added Tax (VAT) data reported later than is needed for monthly estimates [14]. The last two indicators in the Time-related dimension are stability related. The first of these indicators focuses on the dynamics of the population of objects in the current file compared with those in previous deliveries. The second indicator checks if and how the values of a variable (such as the NACE code of a company) change back and forth between subsequent deliveries.

3.5 Integrability

The Integrability dimension contains indicators specific for the ease by which data can be integrated into the statistical production system of an NSI. The effect of problems in the alignment of objects (indicator 5.2) is illustrated in Figure 2. Here, the turnover reported by business units in two distinct administrative data sources is compared. The two sources used are the Intra-Community Performances (ICP) and VAT data of the Dutch Tax and Customs Administration. Both sources contain information on trade

Figure 2. Comparison of turnover reported by units in ICP and VAT data sources



between the Netherlands and other EU-countries. Comparing the trade data in the sources reveals that the export-related turnover data aligns well. The import-related turnover data, however, clearly differs. The fact that export aligns well is caused by the fact that a single Dutch business is responsible for reporting both its VAT and its ICP export data; independent of the number of trading partners. The ICP import data of a Dutch business, however, is reported by its non-Dutch trading partners, which can be many. As a result, the ICP import data is much more difficult to align because it is i) problematic to correctly identify the corresponding unit and ii) it suffers from underreporting.

4. Conclusion

Apart from developing scripts for all measurement methods, WP4 of the BLUE-ETS project will develop a way to report the overall findings in a so-called Quality Report Card (QRC). Goal of the QRC is to present the outcomes of the indicators in an easily readable format at a dimensional level. This very much resembles the presentation of the metadata quality evaluation results as used for the Dutch checklist [5,7]. A draft version of the QRC is included in [9]. During the remainder of the BLUE-ETS project [4], also various administrative data sources -in the various countries involved- will be evaluated with the indicators listed in Table 1. This approach will, for the first time, allow NSI’s to unambiguously determine the input quality of these sources in a structured way.

5. Acknowledgments

The authors wish to acknowledge many individuals whose help and cooperation aided in the completion of the work described in this paper. They are especially grateful to the contributions of Eric Schulte Nordholt, Edwin de Jonge (Statistics Netherlands), Li-Chun Zhang, Coen Hendriks, Kristin Foldal Haugen (Statistics Norway), Fulvia Cerroni, Grazia Di Bella (Italian National Institute for Statistics) Thomas Laitila, Anders Wallgren and Britt Wallgren (Statistics Sweden).

6. References

[1] Arts, C.H., Bakker, B.F.M., van Lith F.J. (2000), Linking administrative registers and household surveys, *Netherlands Official Statistics*, 15, 16-21.

[2] 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, IAOS Conference on Reshaping Official Statistics, Shanghai, China.

[3] Bakker, B.F.M. (2012), Estimating the validity of administrative variables, *Statistica Neerlandica*, 66, 8-17.

[4] BLUE-ETS (2012), Project description on the BLUE-Enterprise and Trade Statistics website, www.blue-ets.eu.

[5] Daas, P.J.H., Ossen, S.J.L. (2011), Metadata Quality Evaluation of Secondary Data Sources. *International Journal for Quality Research*, 5, 57-66.

[6] Daas, P.J.H., Ossen, S.J.L., Tennekes, M. (2010), Determination of Administrative Data Quality: Recent results and new developments, Q2010 European Conference on Quality in Official Statistics, Helsinki, Finland.

[7] 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.

[8] Daas, P., Ossen, S., 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, First deliverable of WP4 of the BLUE-ETS project, March 10.

[9] Daas, P., Ossen, S., Tennekes, M., Zhang, L-C., Hendriks, C., Foldal Haugen, K., Cerroni, F., Di Bella, G., Laitila, T., Wallgren, A., Wallgren, B. (2011), Report on methods preferred for the quality indicators of administrative data sources, Second deliverable of WP4 of the BLUE-ETS project, September 28.

[10] Olson, J.E. (2003), Data Quality: the Accuracy Dimension, Morgen Kaufmann.

[11] Schouten, B., Cobben, F., Bethlehem, J. (2009), Indicators of Representativeness of Survey Nonresponse, *Survey Methodology,* 35, 101-113.

[12] 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, 58th Session of the ISI, Dublin, Ireland.

[13] Tennekes, M., de Jonge, E., Daas, P.J.H. (2011), Visual Profiling of Large Statistical Datasets. New Techniques and Technologies for Statistics conference, Brussels, Belgium.

[14] Vlag, P. (2011), Short Term Turnover Estimates with Incomplete VAT Data, 58th Session of the ISI, Dublin, Ireland.

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

[16] Zhang, L.-C. (2011), A unit-error theory for register-based household statistics, *Journal of Official Statistics,* 14, 415–432.

[17] Zhang, L-C. (2012), Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.