UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

STATISTICAL OFFICE OF THE EUROPEAN UNION

Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses
(The Hague, The Netherlands, 10-11 May 2010)

Working paper 5
28 April 2010

Topic (ii) of the provisional agenda

QUALITY OF REGISTERS

On the quality of registers

Note by the Statistics Netherlands*

Summary: Statistics Netherlands is increasingly making use of administrative registers for the production of statistics. This approach makes Statistics Netherlands not only more dependent on the availability of these types of data sources but also on the quality of those sources. It is therefore of vital importance that a procedure is available to determine, in a systematic, objective, and standardized way, the quality of administrative registers. For this purpose a quality framework was developed. The framework consists of three high level views on the quality of an administrative register. The three hyperdimensions are called: Source, Metadata, and Data. With a checklist the quality aspects included in Source and Metadata, which focus on the exchange and the metadata of the data source, are determined. The study of the quality of the data, the third view in the framework, is not part of the checklist. For data another approach has to be developed. Results of the application of the checklist and ways to study the data quality of registers are discussed in this paper.

Keywords: Quality framework, Administrative data, Register-based statistics

* Prepared by Piet Daas, Methodology sector, Division of Methodology and Quality.
The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

1. Introduction

National Statistical Institutes (NSI's) need data for the production of statistics. Apart from the data obtained through surveys, NSI's are increasingly making use of data collected and maintained by other organizations for non-statistical purposes. Registers and administrative data sources are examples of such data sources (Wallgren and Wallgren, 2007). They are produced as a result of administrative processes of organizations but are -very often- also interesting data sources for NSI's. During the last decade, more and more NSI's have realized this (Unece, 2007). A major advantage of using administrative registers for statistics is the fact that it drastically reduces the costs of data collection and the response burden on enterprises and persons. Since administrative registers often completely cover whole populations, in various time references, they are also particularly well suited for censuses (Schulte Nordholt, 2004) and for creating detailed and longitudinal statistics on (sub)populations and regions (Wallgren and Wallgren, 2007).

From a statistical point of view, administrative registers have some disadvantages. The most important one is the fact that the collection and processing of the data is beyond the control of the NSI. It is the data source keeper who manages these aspects. The same is true for the units and variables a register contains. These are defined by administrative rules and may therefore not be identical to those required by an NSI (Wallgren and Wallgren, 2007). It often takes considerable effort to clearly determine the statistical usability of an administrative register (Bakker et al., 2008). Since the production of high quality statistics depends on the quality of the input data, it is of vital importance that NSI's are able to determine the quality of administrative registers preferably in a cost efficient way. For this purpose a quality framework was developed by Statistics Netherlands that enables the determination of the quality of externally collected, secondary, data sources, such as administrative registers (Daas et al., 2008).

2. Quality framework

At Statistics Netherlands an extensive literature study was performed to identify the various quality aspects of administrative data. This 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). 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- 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 preferred in this paper.
With three hyperdimensions all quality aspects identified in the literature study could be combined into a single framework (Daas et al., 2008). These hyperdimensions are called: Source, Metadata, and Data. Each hyperdimension is composed of several dimensions of quality; 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 (Daas et al., 2008). The paper starts with an overview of the quality aspects in the Source and Metadata hyperdimension and the method developed to determine them. Next, recent insights on the study of the quality aspects in the Data hyperdimension are described.

Source and Metadata hyperdimensions

An NSI that plans to use an administrative register should start by exploring the quality aspects of the information required to enable the use of the data source on a regular basis. These aspects of quality are located in the Source hyperdimension of the quality framework. In table 1 the dimensions, quality indicators, and method descriptions for the Source hyperdimension are listed. The Metadata hyperdimension focuses on the conceptual and process related quality aspects of the metadata of the source. It is essential that an NSI fully understands these metadata related quality aspects because any misunderstanding(s) highly affect the quality of the output based on the data in the source. In table 2 the dimensions, quality indicators, and method descriptions are shown for the Metadata hyperdimension.

Table 1: Quality framework for registers, Source hyperdimension

DIMENSIONS	QUALITY INDICATORS	METHOD DESCRIPTION
1. Supplier	1.1 Contact	- Name of the data source
		- Data source contact information
		- NSI contact person
	1.2 Purpose	- Reason for use of the data source by NSI
2. Relevance	2.1 Usefulness	- Importance of data source for NSI
	2.2 Envisaged use	- Potential statistical use of data source
	2.3 Information demand	- Does the data source satisfy information demand?
	2.4 Response burden	- Effect of data source on response burden
3. Privacy and security	3.1 Legal provision	- Basis for existence of data source
	3.2 Confidentiality	- Does the Personal Data Protection Act apply?
		- Has use of data source been reported by NSI?
	3.3 Security	- Manner in which the data source is send to NSI
		- Are security measures required? (hard/software)
4. Delivery	4.1 Costs	- Costs of using the data source
	4.2 Arrangements	- Are the terms of delivery documented?
		- Frequency of deliveries
	4.3 Punctuality	- How punctual can the data source be delivered?
		- Rate at which exceptions are reported
		- Rate at which data is stored by data source keeper
	4.4 Format	- Formats in which the data can be delivered
	4.5 Selection	- What data can be delivered?
		- Does this comply with the requirements of NSI?
5. Procedures	5.1 Data collection	- Familiarity with the way the data is collected
	5.2 Planned changes	- Familiarity with planned changes of data source
		- Ways to communicate changes to NSI
	5.3 Feedback	- May NSI contact data source keeper 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

2.1.1 Checklist for Source and Metadata

For the evaluation of the quality aspects in the Source and Metadata hyperdimension a checklist has been developed (Daas et al., 2009). It is included in the paper of Daas et al. (2009) which can be downloaded from the Statistics Netherlands website. The checklist guides the user through the measurement methods for each of the quality indicators in both hyperdimensions. By answering the questions in the checklist, the ‘value’ of every method for each indicator in tables 1 and 2 is

determined. Evaluation of the Metadata-part requires that the user has a particular use in mind. To test the usability of the checklist and its usefulness for statistics, eight administrative data sources were evaluated. These data sources 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).

Because our primary interest in this study was the usability of the outcome of the checklist, the checklists were not self-administered but filled in in close cooperation between one (or more) of the authors and several key staff members of our office. The latter were involved in: i) contact with the data source keeper, ii) receipt of the data source, and iii) processing/checking of the data source. An average, it took about 2 hours to complete the checklist.

Table 2: Quality framework for registers, Metadata hyperdimension

DIMENSIONS	QUALITY INDICATORS	METHOD DESCRIPTION
1. Clarity	1.1 Population variable definition	- Clarity score of the definition
	1.2 Classification variable Definition	- Clarity score of the definition
	1.3 Count variable definition	- Clarity score of the definition
	1.4 Time dimensions	- Clarity score of the definition
	1.5 Definition changes	- Familiarity with occurred changes
2. Comparability	2.1 Population variable definition comparison	- Comparability with NSI definition
	2.2 Classification variable definition comparison	- Comparability with NSI definition
	2.3 Count variable definition comparison	- Comparability with NSI definition
	2.4 Time differences	- Comparability with NSI reporting periods
3. Unique keys	3.1 Identification keys	- Presence of unique keys - Comparability with unique keys used by NSI
	3.2 Unique combinations of variables	- Presence of useful combinations of variables
4. Data treatment (by data source keeper)		
	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

2.1.2 Checklist results

The evaluation results obtained for the eight data sources studied are shown in tables 3 and 4. These tables contain the results for the Source and Metadata hyperdimension, respectively. For the IPA, the Metadata 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. For the other data sources, the envisaged use was educational statistics.

Evaluation scores are indicated at the dimension level. The dimensional scores were obtained by selecting the most commonly observed score for every measurement method in each dimension. The symbols for the scores used in tables 3 and 4 are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combing symbols with a slash (/) as a separator. When in a specific dimension an unclear score occurs for a specific quality indicator this score is shown for the whole dimension. If the scores for the other indicators in that dimension are not unclear, the most commonly observed score for these indicators is additionally included between brackets.

The results in table 3 reveal that on a dimensional level, the overall scores for the data sources are somewhat low on the delivery and procedures dimension in Source. For the first dimension this is predominantly caused by the not always timely delivery of the IPA, CWI, and 1FigSGE. This implies a possible risk for users that rely heavily on the timely availability of these data sources. The major problem here is the delivery of the CWI. The CWI is hardly ever delivered on time; a delay of a

Table 3. Checklist results for the Source hyperdimension

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

Table 4. Checklist 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)	?(+)	?(+)	+	+

few weeks is not uncommon. There even has been a period of three months when no data was delivered at all. The somewhat low scores for IPA and 1FigSGE are not unexpected; both data sources have relatively recently been created. Because of this delivery times still fluctuate somewhat.

For the procedures dimension, the scores are somewhat low because of the low scores on the fall-back scenario indicators (see table 2). Not all users were aware of the fact that -in our office- such a scenario does not have to be developed for all administrative data sources. This affected the score in a negative way. With this in mind, hardly any procedural problems were observed, except for the NCP. Here, contact with the data source keeper is somewhat troublesome. Request for (additional) information are not always timely replied and the answers provided are not always very clarifying.

The results for the Metadata hyperdimension are shown in table 4. Compared to the Source hyperdimension (table 3) clearly more poor (-) scores are observed. Here again the CWI attracts attention. This data source scores negative in the clarity and comparability dimensions. For both dimensions, this is largely the result of the discrepancy between the definition of the CWI-variable 'level-of-education' and the definition of the corresponding variable of Statistics Netherlands. Detailed studies revealed that the interpretation of the 'level-of-education' variable at CWI is highly affected by the combination of the study history and discipline of a job-seeker and the jobs available (Bakker et al., 2008). The CWI sometimes down- or upgrades the 'level-of-education' of job-seekers when hardly any or quite some jobs are available in their discipline. This is purely done to increase their change of finding a job. CWI clearly has a more practical, less strict, interpretation of the 'level-of-education' variable than Statistics Netherlands. On a dimensional level, the data treatment dimension is the most unclear area in Metadata. It shows that in our office relatively little knowledge is available on the (possible) checks and modifications of the data by the data source keeper. Positive exceptions to this are the IPA, NCP, and MBA.

Overall the evaluation results for the eight data sources reveal that attention should be paid to the delivery, clarity, and comparability related quality aspects of the CWI before any studies are performed that relate to the data in this source. Only when these problems are solved satisfactory it makes sense to spend (a lot more) time and effort in the determination of the quality of the CWI-data. The results for the MBA demonstrate that it is possible to have every quality aspect in the 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. For all data sources, except the CWI, the next logical next step is the determination of the quality of the data. This is the topic of the next section.

2.2. Data hyperdimension

The quality aspects proposed for the Data hyperdimension are listed in table 5 (Daas et al., 2008). Many of the quality indicators in table 5 are familiar to statisticians, but some are probably not. The latter will be briefly discussed. A considerable part of the measurement methods in Data is based on the so-called Representativity index (R-index). This is an indicator 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. Representative is, however, also important for administrative data because the composition of the units present in such a data source may be time-dependent. In the Netherlands, for instance, the composition of 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 was not added as a separate dimension in the Data hyperdimension. The dimension Precision in table 5 is also mainly used to determine the effect of time-dependent changes in the population composition on data quality.

Table 5. Quality framework for registers, Data hyperdimension

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 [1)] for unit composition
	3.3 Effect on average	-Maximum bias of average for core variable
		-Maximum RMSE [2)] 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

[1] 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). [2] 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.

2.2.1 Structured study of data quality

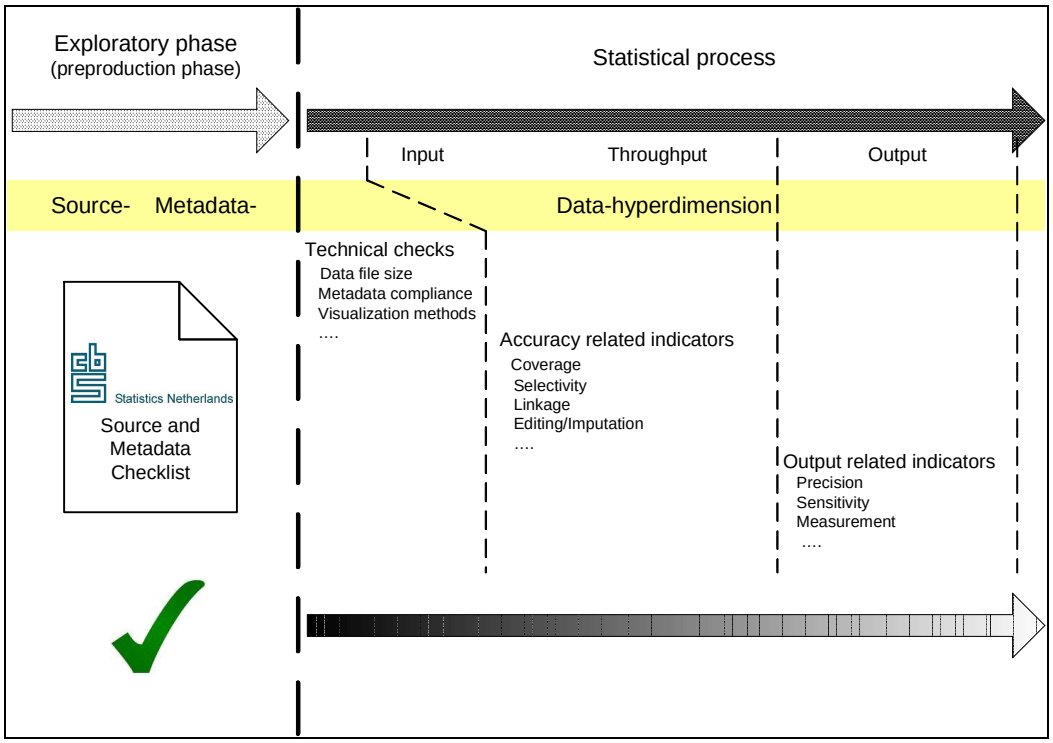
When studying the quality of the data of administrative registers quite some quality aspects have to be determined. The list in table 5 contains a total of 26 measurement methods. It will not be very efficient for an NSI to measure the value of each indicator every time a data source is received. Let 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 indicators should be determined first. When problems are observed at that stage, more detailed studies can be performed.

An overview of the various stages discerned in Data is shown in figure 1. For completeness, this figure also contains the checklist for the quality aspects in the Source and Metadata hyperdimension. The most important part of figure 1 is the subdivision of the study of data quality in three stages. These stages differ greatly in their level of effort and detail. The stages are: i) Technical checks, ii) Accuracy related quality indicators, and iii) Output related quality indicators. They are discussed below.

2.2.2 Stages of data quality

In the technical checks stage, quick checks are performed of the data delivered to the NSI. The checks are very basic, need to identify serious errors, and need to be performed quickly. Examples of technical checks are the comparison of file size and number of (unique) units delivered and compliance of the data to the metadata. An interesting addition to this are visualization based checks; an approach already used considerably in the data exploratory phase of large data sources (Uwin et al., 2006).

Figure 1. Overview of the various stages in quality aspect measurement



When an NSI has decided for which publication a data source is going to be used, more specific quality indicators can be applied. The indicators belong to the second stage. They 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 5 belong to this group. Examples of indicators for units are: over- and undercoverage, 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.

The quality indicators in the third stage are all output oriented. They report on a level that, for an NSI, ultimately determines the statistical usability of a data source; that of the quality of the output. Examples of indicators in this group are indicators that aim to determine the precision of core variables and the selectivity of composite totals.

There is however a restriction to the indicators at the third stage, and probably also to some of those included in the second stage, discerned. These types of quality indicators 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). 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 Data hyperdimension has to be restricted to some extent as it is impossible to meet all conceivable uses of every user.

2.2.3 Implementation of data quality determination

Studies of the quality indicators in the Source and Metadata hyperdimension have shown that implementation is greatly enhanced when it is embedded in a structured approach; i.e. the checklist (see above). A similar approach should be developed for data quality. In addition, availability of standardized script and/or a software implementation of (a large part of) the checks and indicators in the Data hyperdimension will speed of the measurement process and greatly assist the user. At the end of this process, all the quality scores for Data should be combined into a single instrument. This to assure the central availability of the quality results for tha data sources studied.

3. Concluding remarks

The results described in this paper show that the quality framework developed for administrative registers and the corresponding checklist are valuable tools for the evaluation of the statistical usability of such data sources. Advantage of the use of the checklist is that it: i) provides a structured way of looking at the Source and Metadata quality aspects and that ii) not immediately a great deal of attention and work is put into data related quality aspects. The latter is often the case in practice. When the checklist does not reveal any serious problems for a data source, the quality aspects included in the Data hyperdimension should be determined. The latter hyperdimension is the focus of current research. Main topics being studied are the development of a structured approach for efficiently evaluating the quite large number of quality indicators in this hyperdimension, the use of visualization methods, and the use of standardized scripts or software tools to speed up the determination process and assist the user.

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. In: 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., Vis-Visschers, R.J.W.M., Arends-Toth, J. (2009) 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

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), pp. 101-113.

Schulte Nordholt, E.(2004) Introduction to the Dutch virtual census of 2001. In: The Dutch Virtual Census of 2001, analysis and methodology, Eds., Schulte Nordholt, E., Hartgers, M., Gircour, R., Statistics Netherlands, Voorburg.

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.

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

*****