

EUROPEAN COMMISSION
RESEARCH DIRECTORATE-GENERAL



BLUE-Enterprise and Trade Statistics
BLUE-ETS

SP1-Cooperation-Collaborative Project
Small or medium-scale focused research project

FP7-SSH-2009-A

Grant Agreement Number 244767

SSH-CT-2010-244767

Deliverable WP8.2

Title: Guidelines on the usage of the prototype of the computerized version of QRCA, and Report on the overall evaluation results.

Authors: Piet Daas, Martijn Tennekes, Saskia Ossen (CBS),
Grazia Di Bella, Lorena Galiè (ISTAT),
Thomas Laitila, Daniel Lennartsson, Richard Nilsson, Anders Wallgren,
and Britt Wallgren (SCB),

With contributions from: Joep Burger (CBS), Daniela Bonardo, Fulvia Cerroni,
Valentina Talucci (ISTAT), Coen Hendriks, Li-Chun Zhang and Kristin Foldal
Haugen (SSB).

DRAFT

DATE 29-03-2013

Deliverable WP8.2

Guidelines on the usage of the prototype of the computerized version of QRCA, and Report on the overall evaluation results

Summary:

In this report the findings of the evaluation of various administrative data sources used by Dutch, Italian and Swedish National Statistical Institutes are reported. Administrative data sources were studied from both the data source quality and input oriented output quality point of view (deliverable 4.2). Since not all of the institutes were able to directly apply the statistical measurement methods implemented in the R-package developed in WP 4 (deliverable 4.3), the measurement methods required were implemented in the (standard) programming language of choice for each institute. The experiences and results obtained by the NSIs involved are generally discussed in this report. In the appendices, the findings of each source for each individual NSI are included. This report also contains the manual for the methods implemented in the dataquality R-package and the most recent version of the Quality Report Card for Administrative data (QRCA).

Contents

1. Introduction.....	4
2. Data Quality Evaluations	5
2.1. Data sources and approach used	5
2.1.1. Statistics Netherlands.....	5
2.1.2. Italian National Institute of Statistics.....	5
2.1.3. Statistics Sweden.....	6
2.2. Evaluation findings	6
2.2.1. Statistics Netherlands.....	6
2.2.2. Italian National Institute of Statistics.....	7
2.2.3. Statistics Sweden.....	7
3. Overall conclusions.....	8
References.....	9
Appendix A: Evaluation results of Statistics Netherlands.....	11
Appendix B: Evaluation results of Italian Institute of National Statistics	42
Appendix C: Evaluation results of Statistics Sweden.....	119
Appendix D: Manual for determining the input quality of administrative data sources with the ‘dataquality’ package including the Quality Report Card	150

1. Introduction

Many National Statistical Institutes (NSIs) 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 (UNECE, 2007). The preconditions that enable a NSI to extensively make use of administrative sources in statistics production are: 1) legal foundation for the use of 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 those sources, it is essential that a 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 a NSI. It is the data source holder that manages these aspects. It is therefore of vital importance that a 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 standardized way.

For the evaluation of the *metadata* quality components of administrative data sources a procedure (Daas *et al.*, 2009) has already been developed. This approach has been thoroughly evaluated in the Netherlands and proved very useful (Daas and Ossen, 2011). 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 (Daas *et al.*, 2010). This view is commonly referred to as the input quality of administrative data (Daas *et al.*, 2011a-b; 2012a).

The development of an approach that can be used to routinely determine the input quality of administrative data -for statistical purposes- has been the main focus of Workpackage 4 (WP4) of the BLUE-Enterprise and Trade Statistics (BLUE-ETS) project (BLUE-ETS, 2012). In addition, the need to develop indicators and measurement methods to evaluate administrative data sources for potential use in statistics production emerged (BLUE-ETS, 2011b; Laitila, 2012;). Clearly, the quality of administrative data as it enters the office and the effect the quality of this data on the statistics produced need both be considered. As a result, the latter work was additionally studied (Laitila *et al.*, 2011, 2012).

The final products of Workpackage 4 of the BLUE-ETS project were:

- A) Methods to determine the input quality of administrative data sources (DSQ; Del. 4.2 & 4.3)
- B) Methods to evaluate the quality of administrative sources for potential use in statistics production (IOQ; Del. 4.2 & 4.3)
- C) Quality Report Cards for both views (Del. 4.3)
- D) A first version of an R-package in which methods to determine the input quality of administrative data sources are implemented (addition to Del. 4.3)

This report describes the results of the application of the above mentioned products on various administrative data sources at the NSIs involved and is organized as follows. In chapter 2 an overview is given of the approach used and the data sources evaluated. Chapter 3 discusses the overall findings and general conclusions. The detailed findings of each NSI are included as appendices to this report. In addition, the manual for the use of the methods implemented in the dataquality R-package and a finalized version of the Quality Report Card for Administrative data

(QRCA) are added. The dataquality package will soon be published on the Comprehensive R Archive Network. Meanwhile, the most recent version of the ‘dataquality’ R-package can be obtained by contacting one of the Dutch (CBS) authors of this paper. With this report the study of the quality of administrative data sources from a statistical point of view by the BLUE-ETS project is finalized.

2. Data Quality Evaluations

In this chapter an overview is given of the sources evaluated, the viewpoint used and the way the measurement methods identified in deliverable 4.2 (Daas *et al.*, 2011b) were implemented. The various NSIs involved in this part of WP8 each selected one or more data sources and evaluated its quality from the viewpoint that best suited their needs. During WP4 of the BLUE-ETS project it became more and more clear that each NSI has its own particular way of looking at the quality of administrative *data* (Daas *et al.*, 2012b; Laitila, 2012). These views are heavily influenced by the data quality ‘culture’ at the NSI and the possibilities (limitations) the IT-infrastructure offers. The former is reflected in the different viewpoints on the input quality of administrative data, e.g. DSQ vs. IOQ (Daas *et al.*, 2012b). The latter is greatly affected by the possibility to use the methods implemented in the R-package developed as part of deliverable 4.3 (Daas *et al.*, 2012b).

By providing each NSI with the opportunity to determine the quality of administrative data from their own viewpoint and choose their own way of implementing the measurement methods proposed, it was assured that each NSI *was* actually able to evaluate administrative data sources and obtain results they could use. The need for such a flexible approach with the few NSIs involved, demonstrates the diversity in the quality checking environment at each NSI and point to the need for various ways to implement the theoretical framework developed in WP4. As a result of this choice it became apparent that only at Statistics Netherlands could the R-script in the dataquality package be applied directly. At the other NSIs involved, additions to R (e.g. packages) could not be easily included. These NSIs decided to implement the measurement methods described in deliverable 4.2 in their own language of choice; such as SAS (see Appendix B and Laitila *et al.*, 2012). This observation also reduced the need for a graphical interface for the R-scripts (Daas *et al.*, 2012b); this work was subsequently halted.

2.1. Data sources and approach used

For each participating country the data sources studied, the viewpoint used and the way the measurements were implemented are described below.

2.1.1. Statistics Netherlands

At Statistics Netherlands two administrative data sources were studied that are used for the production of enterprise statistics: Value Added Tax (VAT) data and the Insurance Policy record Administration (IPA). In the Netherlands, quarterly reported VAT-data received from the Dutch tax-office are combined with survey data for the largest companies. The combined source is subsequently used to -for instance- produce short term statistics. IPA data is maintained by the Dutch Institute for Employee Benefit Schemes and contains data provided by businesses. The IPA contains very detailed information on jobs, pensions and benefits. The data source is becoming increasingly important for many of the enterprise and social statistics in the Netherlands. Both data sources were studied from the input quality point of view (DSQ). The methods implemented in the dataquality R-package (version 3.6) were used. More details are included in Appendix A.

2.1.2. Italian National Institute of Statistics

The Italian National Institute of Statistics (ISTAT) studied Social Security Data (SSD). This data source is produced by INPS, the Italian Institute of Social Security, and concerns the monthly

contribution declaration of employers for employees. The SSD has been recently acquired for statistics production at ISTAT and is interesting for both enterprise and for social statistics. The viewpoint from which this data source was studied is the input quality point of view (DSQ). The methods developed in WP4 were implemented in the statistical software SAS (Statistical Analysis System). More details are included in Appendix B.

2.1.3. *Statistics Sweden*

As a result of the need to develop a procedure to evaluate administrative data sources for potential use in statistics production (Daas *et al.*, 2011b), our Swedish colleagues studied the quality of three data sources according to the evaluation system described in Laitila *et al.* (2012). This broad perspective enables thorough evaluation of administrative data (registers) for statistics production. The source studied are: the Income Statement register, the VAT-register and the Annual Company Reports (SRU) register, respectively. The Income Statement register contains data on yearly wages and salaries, preliminary tax and benefits regarding each employee. About 60 % of total taxes in Sweden are covered by this system. The VAT-register data is obtained from several sources, i) the monthly reported VAT by businesses to the Tax Board, ii) VAT reported to the Tax Board in the income-tax returns of smaller businesses and iii) information on VAT retrieved from the Customs office every six months. All data is delivered to Statistics Sweden. The SRU-register contains the yearly tax returns for Sole proprietorships, Trading partnerships, Limited partnerships, Limited companies and Economic associations. They consist of three parts: balance sheet, profit and loss statement and tax adjustments. Limited companies provide more information and sole proprietors provide less detailed information. The SRU is the main source for the Structural Business Statistics. More details are included in Appendix C.

2.2. Evaluation findings

For each data source as many of the measurement methods implemented in the scripting language of choice were evaluated. However, in some cases, for instance when the required information was not included in the dataset or a reference population was not (yet) available, automated evaluation was not possible. In such cases, data source experts were contacted in an attempt to provide the information needed. Results obtained were noted on Quality Report Cards and described in reports. Both are included as appendixes to this document. Below a general overview is provided of the most important findings for each NSI who participated in these studies.

2.2.1. *Statistics Netherlands*

The evaluation findings of the two data sources studied produced a remarkable outcome. The quality of the data in both sources suffered from some serious issues. The VAT-data evaluation revealed a serious completeness issue (first report card in Appendix A). For some companies turnover data was missing. Depending on the opinion of a SN-expert and the size class of the companies involved, additional data may be needed to solve this issue. For the IPA an accuracy issue was found (second report card in Appendix A). It was found that the data source contained three persons that worked for 52 or more companies. Again a data source expert needed to be consulted to reveal the seriousness of this issue. It is not the outcome of the experts that interests us here, it is the fact that a structured evaluation of administrative data immediately revealed an issue in the both the sources studied that surprised us. Both sources demonstrate the advantages of determining as many indicators as possible, preferably in various ways, for administrative data. Fully automating such an approach could, for instance, enable evaluation overnight of each delivery and highlighting any issues. Downside of such an approach is that it, potentially, could take up considerable time and computational resources if applied to each delivery of a large data source (Daas *et al.*, 2010).

The results for the technical dimension (added to the end of each QRCA, see Appendix D) already indicated some of the issues confirmed later on. It is interesting to meticulously look at the results

of the described function added for each data source at the end of the report cards of Appendix A. The studies also showed that both relatively small (VAT) and large (IPA) data sources could be evaluated with the R-scripts. The visualization method, the tableplot, developed in WP4 (Tennekes *et al.*, 2011, 2013) was also very insightful.

Writing down the evaluation results in QRCA demonstrated the advantage of noting quality findings in a structured way. It was questioned, however, if the structure induced by the sequence in which the dimensions are listed on the QRCA sufficed. The ISTAT findings (see below) even more clearly confirmed this. In essence, only the first dimension needs to be valued first. For the other dimensions, there is no real reason to strictly follow the sequence used on the form or proposed in the theoretical framework (Daas *et al.*, 2012b). What also became apparent was the need to add additional findings (such as plots) to the report card. This was done in the QRCA's of Appendix A and has been included in the new version of the QRCA (see Appendix D).

2.2.2. *Italian National Institute of Statistics*

The SSD source reacted well to the tests. This demonstrated that the measurement methods were successfully implemented in SAS and could be easily applied. In SAS huge amounts of data could be evaluated in the ISTAT IT-environment. All quality indicators studied reached the objectives and showed some interesting findings (Appendix B) but no serious issues were found.

The most important findings were that the implementation of the Quality Report Card lacked the highlighting of the strict interrelation between dimensions. It could be useful to include these. The downside is that these links could negatively affect the layout and 'ease of use' of the report card. For this reason alone, interrelation was not added. The ISTAT findings also suggested defining a new hierarchical chain for the indicators in the dimension identified. Such a chain would more clearly reflect the way the evaluation proceeds in practice. The suggested chain is Technical checks, Integrability, Accuracy, Completeness, Time-related dimension. This suggestion has been included in the new version of the QRCA (Appendix D)

2.2.3. *Statistics Sweden*

In Sweden three data sources were evaluated by the approach described in Laitila *et al.* (2012). The indicators are grouped in sets which provides a working procedure involving the sequence of evaluating i) metadata, ii) accuracy, iii) integration with a base register, and iv) integration with other surveys. The definitions of the indicators are suggested with the view of implementing a data source (register) within statistical register systems as proposed by Wallgren and Wallgren (2007).

It is natural to first consider the (metadata) contents of an administrative data source and then evaluate the accuracy of the contents. For integration it is of interest to find out if the source can be incorporated into the system by relating it to a base register. Performing these three first steps of the evaluation process provides insights as to the sources themselves and other data sources used in the evaluation process. This is perhaps one of the more important results in the applications reported and an issue for future projects. To maintain a fully functional register system requires a continuous evaluation of the consistency with relevant external information. Here any kind of relevant information is of interest and not only administrative sources potentially useful for statistics production.

The fourth set of indicators addresses the issue of how an administrative data source can be utilized for improving surveys conducted at an NSI. Again the experience of the applications is that this step provides both information on the quality of the source evaluated and the data sources used for the evaluation. It was also found that this evaluation is a methodologically demanding task. This step will also involve subjective judgments since the considered potential usage of the new data sources depends on the experiences of those performing the evaluation.

3. Overall conclusions

The most important finding is obvious from the previous chapter and the introduction. Evaluation of the quality of administrative data sources from a statistical point of view is difficult and demanding in practice. Not only is this caused by the two clearly distinguishable points of view (DSQ and IOQ), the actual evaluation of sources is also severely affected by the IT-environment of an NSI and hence the software available. This prompted us to become (even) more flexible in our plans and focus less on the programming language in which the measurement methods were evaluated. This observation is certainly one that is important for the future use of the evaluation approach developed at other National Statistical Institutes.

Because of the various ways in which the measurement methods were evaluated, there was less need for a single ‘package’ in which the methods were implemented. The R-package initially created as part of WP4, however, still had an important function. First of all, it demonstrated which methods could actually be implemented and which could not (Daas *et al.*, 2012). Secondly, it assured that the methods that could be implemented were thoroughly checked. By virtue of the package and the synthetic data included, it also enables demonstrations of the methods and the advantages of the quality evaluation approach developed in WP4. For these reasons the dataquality R-package will be made officially available in due time. Prior to publishing the package on the Comprehensive R Archive Network, the most recent version of the ‘dataquality’ R-package can be easily obtained by contacting the Dutch (CBS) authors of this paper.

The QRCA demonstrated its use as a way of structurally noting the DSQ-findings. The sequence used, however, needed to be changed. As a result -in accordance with the suggestion of ISTAT- the original sequence of the dimensions has been altered to: Technical checks, Integrability, Accuracy, Completeness, Time-related dimension. The need to add additional findings (such as plots) to the report card also emerged. An adjusted version of the QCRA (version 0.4) in which both changes have been included can be found in Appendix D.

A very thorough evaluation of the quality of administrative data sources was applied in Sweden. Here the natural sequence of steps was rigorously implemented. As a result, a quite demanding but highly systematic evaluation process was developed which revealed insights on the quality of the sources themselves and on the other sources used in the evaluation process. Again results were noted in a QRCA to structurally note the evaluation findings.

The time required to thoroughly evaluate administrative data is something that is a serious issue for many NSI’s. It is difficult to conceive of an NSI that routinely checks each quality aspect of the data for *every* delivery of an administrative source. Budget restraints would simply prevent this. Apart from automated checking of some aspects, which is -for instance- routinely done for a range of administrative data sources in the Netherlands, the visualization approach developed in WP4 of the BLUE-ETS project provides a reasonable alternative. It is fairly easy to create a tableplot of a data source and plot the ‘profiles’ of a selected number of variables. Looking and comparing such data ‘pictures’ for subsequent deliveries does not have to take much time and can be used to quickly identify any important data quality issues; see the paper of Tennekes *et al.* (2013) for more details. A totally alternative option is to increase co-operation with the data source holder who could, for instance, implement additional checks upon request of the NSI. Statistics Norway is pursuing this approach (Hendriks, 2012). The future will tell which approach works best.

From the above it has become clear that NSIs need to be aware of the fact they are not in control of the collection of secondary data and, as a result, need to control the quality of this data as it enters the office. The WP4 and WP8 work of the BLUE-ETS project have provided several ways to achieve this. Which way is best? The choice is up to you!

References

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

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.

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.

Daas, P.J.H., Ossen, S.J.L. (2011), Metadata Quality Evaluation of Secondary Data Sources. *International Journal for Quality Research*, 5 (2), 57-66.

Daas, P.J.H., Ossen, S.J.L., Tennekes, M., Burger, J. (2012a) Evaluation and visualisation of the quality of administrative sources used for statistics. Q2012 European Conference on Quality in Official Statistics, Athens, Greece.

Daas, P., Ossen, S., Tennekes, M., Zhang, L-C., Hendriks, C., Foldal Haugen, K., Bernardi, A., Cerroni, F., Laitila, T., Wallgren, A., Wallgren, B. (2011a), List of quality groups and indicators identified for administrative data sources, First deliverable of WP4 of the BLUE-ETS project, March 10.

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. (2011b) Report on methods preferred for the quality indicators of administrative data sources. Second deliverable of workpackage 4 of the BLUE Enterprise and Trade Statistics project, September 28.

Daas, P., Ossen, S., Tennekes, M., Zhang, L-C., Hendriks, C., Foldal Haugen, K., Laitila, T., Wallgren, A., Wallgren, B., Bernardi, A., Cerroni, F. (2012b) Quality Report Card for Administrative data sources including guidelines and prototype of an automated version. Third deliverable of workpackage 4 of the BLUE Enterprise and Trade Statistics project, June 30, version 1.0 draft

Hendriks, C. (2012) Input Data Quality in Register-Based Statistics: The Norwegian Experience. Paper for the Joint Statistical Meeting, San Diego, U.S.A.

Laitila, T. (2012) Quality of registers and accuracy of register statistics, Paper and presentation for the European Conference on Quality in Official Statistics 2012, Athens, Greece

Laitila, T., Wallgren, A., Wallgren, B. (2011) Quality Assessment of Administrative Data. Paper for the 2011 European NTTS conference, Brussels, Belgium.

Laitila, T., Wallgren, A., Wallgren, B. (2012) Quality Assessment of Administrative Data Data Source Quality. Part of the third deliverable of workpackage 4 of the BLUE Enterprise and Trade Statistics project, June 8.

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

Tennekes, M., de Jonge, E., Daas, P.J.H. (2013) Visualizing and Inspecting Large Datasets with Tableplots. *Journal of Data Science*, 11 (1), 43-58.

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

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

Appendix A: Evaluation results of Statistics Netherlands



Quality Report Card for Administrative data



Version v0.3

BLUE-ETS project

Data source studied

	Quarterly reported VATA-data (combined with survey data for large companies)
--	--

Document Version

Version	Adaptations	Responsible	Date
0.2	Based on the initial idea and the proposal included in BLUE-ETS WP4 deliverable 2 a Quality Report Card for Administrative data is proposed	Statistics Netherlands	22/06/2012
0.3	Adjusted version as result of WP8 case studies	Statistics Netherlands	18/01/2013

Document description

This document includes a Quality report Card for Administrative data used for the evaluation of the statistical usefulness of administrative data. The list focuses on the essential quality dimensions and indicators identified in deliverables 1 and 2 of BLUE-ETS workpackage 4. Evaluation is performed by a user at a National Statistical Institute (NSI). The outcome of the evaluation assists the user in the decision to use the data for statistics production.

Instructions

To report the evaluation findings of the quality indicators identified for administrative data a quality report card has been developed. It is named a Quality Report Card for Administrative data (QRCA) and is used to standardize reporting of the evaluation results obtained. Both the findings of automated (scripted) and non-automated quality indicators methods can be noted in the QRCA.

The users starts filling in the QRCA by reporting the results for the Technical Checks dimension, followed by the Time-related, Completeness, Accuracy and Integrability dimension findings. For each dimension, the results of all the measurements methods that can be performed have to be noted down. Quantitative scores can be noted directly. Qualitative score need to be expressed by the signs +, o, - and ? which are used to identify good, reasonable, poor and unclear. Intermediary scores are created by combing symbols with a slash (/) as a separator. This is identical the scores used in the metadata-checklist of Daas and Ossen (2011). Additional space is included to write down remarks.

For each dimension, the evaluation ends with an overall conclusion. This needs to be filled-in by the user. Possible dimensional findings are ‘No problems found’ (green checkbox), ‘Some minor issues observed’ (orange checkbox) and ‘Serious problems detected’(red checkbox). When serious problems are found, this should be noted and –very likely- evaluation must stop. Any instructions included should be followed. When minor issues are found, this also need to be noted but evaluation can continue. When no problems are found evaluation should also continue of course.

In case of serious issues or when all dimensions have been evaluated, the general findings section should be filled in. First, the scores found at the dimensional level are copied and (if needed) converted to the signs +, o, - and ? (see above). If needed, the user can provide additional information at the object or variable level for each dimension. Here also, additional space is included to write down remarks. The summary ends with an overall conclusion for the quality of the data studied which has to be filled in by the user. Plots or other graphical representations of findings can be added to the document if needed.

Information of the NSI-representative who fills in the report

Full name	Piet Daas, Martijn Tennekes, Saskia Ossen
Position	Methodologists
Department	Methodology and Quality
Phone number	+31- <i>not shown for privacy reasons--</i>
E-mail address	< <i>not_shown</i> >@cbs.nl
Date on which the report card was completed	20 December 2012

1. Technical checks

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Score	Remarks
1.1	Readability		+	Data stored in a SQL-database was used
1.2	File declaration compliance		+	All correct
1.3	Convertability		+	SQL-data was exported per sector and stored in native R-format prior to analysis.

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
Data was split at the sector level. This resulted in 6 sector files. The sector Car Trade was studied in detail.	No problems found	<input checked="" type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here: A tableplot was created to additional check the dataset studied (figure 1). The visualise (figure 2) and describe (table 1) scripts were also used to provide an overview of the data. See appendix.		

- Only continue when the GREEN or ORANGE marked area is checked.
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted and the section or person responsible for receipt of the data needs to be contacted.

2. Time-related dimension

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Measurement method used	Score	Remarks
2.1	Timeliness		Contacted expert	o	Info not in dataset. Script could not be applied.
2.2	Punctuality		Contacted expert	o	Info not in dataset. Script could not be applied
2.3	Overall time lag		Contacted expert	o	Info not in dataset. Script could not be applied
2.4	Delay		Contacted expert	o/+	Info not in dataset. Script could not be applied
2.5	Dynamics of objects	Objects	birth, death1, death2, change, change_fast (R-scripts)	b: 10.2% d1: 8.7% d2: 8.6% c: 1.6% cf: 1.6%	Compared unique identifiers <i>Q3 2010 final</i> with <i>Q3 2011 final</i>
2.6	Stability of variables	Variables	changed_value, unchanged_value (R-scripts)	c_a: 5.5% c_b: 28% u_a: 94% u_b: 72%	Compared <i>Q3 2010 final</i> with <i>Q3 2011 final</i> for imputed turnover (a) and number of employees (b)

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
Some scripts could not be tested on the dataset as the dates (of submission and receipt) were not included.	No problems found	<input checked="" type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here: Additional information on the time-related issues were obtained from the Statistics Netherlands expert on this topic. When comparing objects the company identification variable 'beid' was used.		

- Only continue when the GREEN or ORANGE marked area is checked.



When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.

- *When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.*

3. Completeness

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Measurement method used	Score	Remarks
3.1	Undercoverage	Objects	undercoverage (R-script)	u: 8.7%	Compared Q3 2010 final with Q3 2011 final
3.2	Overcoverage	Objects	overcoverage (R-script)	o: 10.2%	Compared Q3 2010 final with Q3 2011 final
3.3	Selectivity	Objects	tabplot (R-package)	+	See tableplot (figure 1)
3.4	Redundancy	Objects	redundancy (R-script)	r_a: 96% r_b: 0%	Compared complete dataset for beid (a) and Q3 2011 final for beid (b)
3.5	Missing values	Variables	redundancy (R-script)	r_a: 2.1% r_b: 95%	For Q3 2011 final on turnover (a) and imputed turnover column (b)
3.6	Imputed values	Variables	redundancy (R-script)	r: 5.2%	For Q3 2011 final on imputed turnover column. % values is % imputed.

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
Not all companies reported turnover for the period studied. The expert has been informed.	No problems found	<input type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input checked="" type="checkbox"/>
Write additional remarks here: Since this data source included a specific variable providing info on imputed values, indicator 3.6 could be easily determined. Despite the serious problem found, evaluation was continued to test the checklist.		

- Only continue when the GREEN or ORANGE marked area is checked.

When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.



- *When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.*

4. Accuracy

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Measurement method used	Score	Remarks
4.1	Authenticity	Objects	checkFormat, changed_value	c_a: 4.9% had x added c_b: 0.6%	For <i>Q3 2011 final</i> on NACE-code (a). Part of codes contained an additional x. On <i>Q3 2010 final</i> and <i>Q3 2011 final</i> for NACE-code value (b)
4.2	Inconsistent objects	Objects	relation	+ (no issues)	For <i>Q3 2011 final</i> NACE-code, Code, beid and 'kernel_id' were compared
4.3	Dubious objects	Objects	relation	+ (no issues)	For <i>Q3 2011 final</i> NACE-code, Code, beid and 'kernel_id' were compared
4.4	Measurement error	Variables	changed_value	c: 56%	Checked beid's reporting turnover in <i>Q3 2011 flash</i> for changed turnover value in <i>Q3 2011 final</i>
4.5	Inconsistent values	Variables	edit_rules	e: 10.4%	Checked consistency between size class and number of employees defined and reported for all data
4.6	Dubious values	Variables	edit_rules	e: 9.1%	Checked for <i>Q3 2011 final</i> if beid's in size class zero all report zero turnover

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
Not all companies in size class zero reported a zero turnover. This is reported to the expert.	No problems found	<input type="checkbox"/>
	Some minor issues observed	<input checked="" type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
<p>Write additional remarks here: Creative use of the script created enable to determine the percentage of beid's that changed their amount of turnover reported between the flash estimate and the final estimate. Suggesting that 56% of the beid's first reported an not completely correct value. The amount of these changes needs to be determined.</p>		

- *Only continue when the GREEN or ORANGE marked area is checked.*
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- *When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.*

5. Integrability

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Measurement method used	Score	Remarks
5.1	Comparability of objects	Objects	coverage	c_a: 91% c_b: 100%	Compared beid's in Q3 2010 final with Q3 2011 final (a) and Q3 2011 flash with Q3 2011 final
5.2	Alignment of objects	Objects	coverage	c_a: 91% c_b: 100%	Compared beid's in Q3 2010 final with Q3 2011 final (a) and Q3 2011 flash with Q3 2011 final
5.3	Linking variable	Variables	Redundancy, checkFormat	r: 0% c: 100%	Checked beid in Q3 2011 final
5.4	Comparability of variables	Variables	changed_value, unchanged_value	c_a: 0.6% c_b: 0% u_a: 99% u_b: 100%	Checked NACE-code changes for beid's in Q3 2010 final and Q3 2011 final (a) and in Q3 2011 flash and Q3 2011 final (b)

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
Relative small changes between subsequent years and no changed in quarter.	No problems found	<input checked="" type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here: Indicator 5.1 and 5.2 are identical because the scripts used beid's. Expert knowledge and additional id-variables are needed to differentiate between both indicators.		

- Only continue when the GREEN or ORANGE marked area is checked.
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.

General findings

Data source studied

Quarterly reported VATA-data (combined with survey data for large companies)
--

Dimensional scores

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Data dimensions	Level	Score	Remarks
1	Technical Checks	Overall	+	
2	Time related	Overall	o	Expert info was used
		Objects	+	
		Variables	+/o	
3	Completeness	Overall	-	Expert contacted
		Objects	+	
		Variables	-	Turnover was missing for some companies
4	Accuracy	Overall	o	Expert contacted
		Objects	+	
		Variables	o	Not all companies that should have zero turnover reported this
5	Integrability	Overall	+	
		Objects	+	
		Variables	+	

Overall conclusion	Overall score	
<i>Write additional remarks here:</i> Missing turnover issue needs to be resolved.	Negative	<input checked="" type="checkbox"/>
	Neutral	<input type="checkbox"/>
	Positive	<input type="checkbox"/>

Appendix

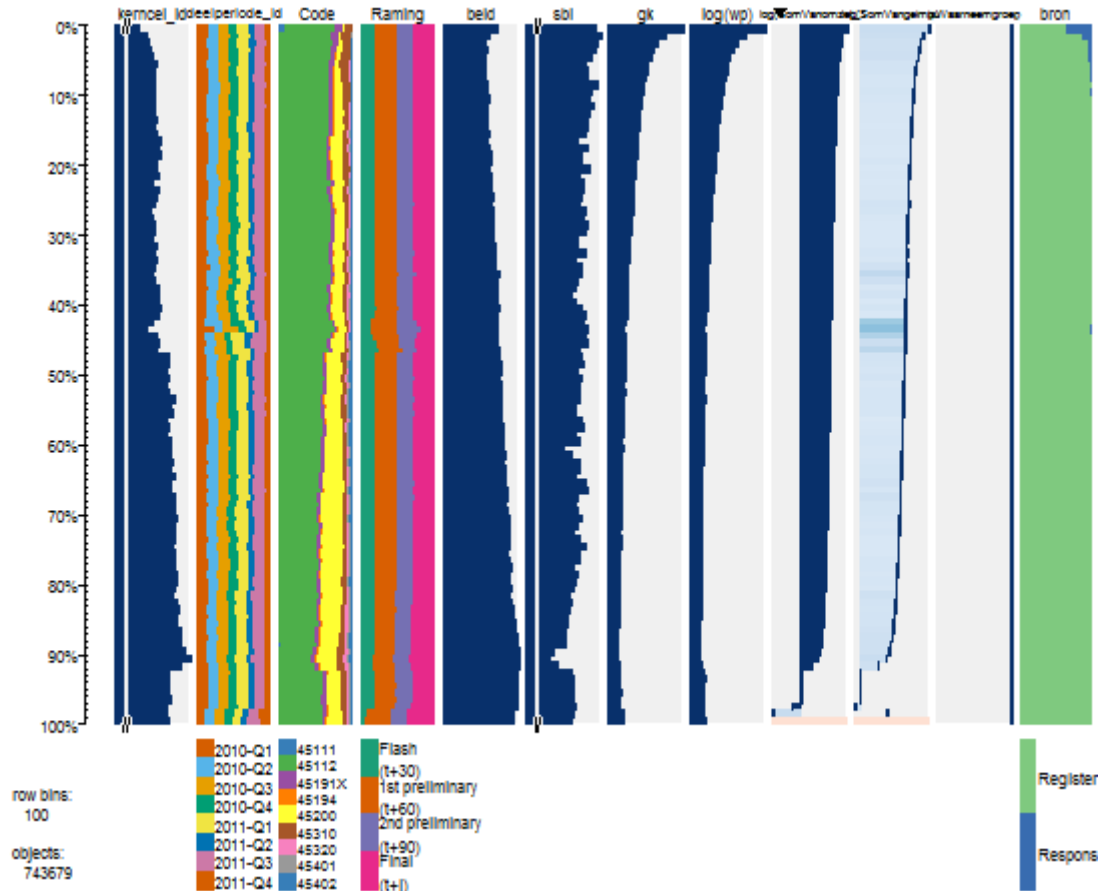


Figure 1: Tableplot for the Car trade sector for the data source studied. Data is sorted on turnover (“SomVanomzet_excl_btwt”). More detailed information on the findings revealed by tableplots are described in the NTS-paper “On the exploration of high cardinality categorical data” of Tennekes and de Jonge (2013).

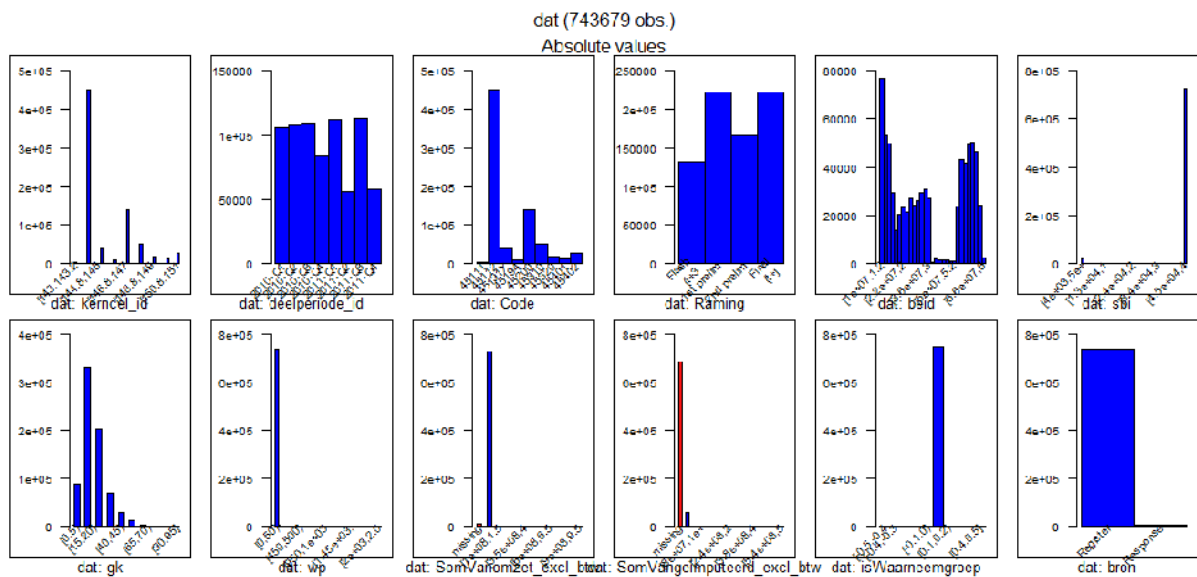


Figure 2: Result of the visualize script for the Car trade sector for the data source studied.

Table 1. Output of the describe script for the Car trade sector for the data source studied.

```

12 Variables      743679 Observations
-----
kerncel_id
      n missing  unique   Mean
743679      0      9  145.4

      143   144   145   146   147   148   149   150   151
Frequency 2272 449873 37586 8431 138958 50297 17307 11316 27639
%          0    60    5    1    19    7    2    2    4
-----
deelperiode_id
      n missing  unique
743679      0      8

      2010-Q1 2010-Q2 2010-Q3 2010-Q4 2011-Q1 2011-Q2 2011-Q3 2011-Q4
Frequency 106061 107480 108712 83411 111816 55858 112692 57649
%          14    14    15    11    15    8    15    8
-----
Code
      n missing  unique
743679      0      9

      45111 45112 45191X 45194 45200 45310 45320 45401 45402
Frequency 2272 449873 37586 8431 138958 50297 17307 11316 27639
%          0    60    5    1    19    7    2    2    4
-----
Raming
      n missing  unique
743679      0      4

Flash
(t+30) (132113, 18%), 1st preliminary
(t+60) (222777, 30%), 2nd preliminary
(t+90) (166022, 22%)

Final
(t+j) (222767, 30%)
-----

```



```

beid
      n missing  unique   Mean   .05   .10   .25   .50   .75   .90
.95
  743679      0   32638 36172029 10865225 11934433 16324110 32442254 57255210 61952060
63509563
  
```

lowest : 10000062 10000119 10000127 10000305 10000534, highest: 66790492 66790530 66791200 66791359 66792657

```

-----
sbi
      n missing  unique   Mean   .05   .10   .25   .50   .75   .90   .95
743679      0     16  44218  45112  45112  45112  45112  45204  45311  45401

      4532 45111  45112 45191 45192 45193 45194 45201 45202 45203 45204 45205 45311 45312
45401 45402
Frequency 17307  2272 449873  1842 27658  8086  8431  3389 10729 17244 51153 56443 43559  6738
11316 27639
%         2     0    60     0    4     1     1     0     1     2     7     8     6     1
2     4
  
```

```

-----
gk
      n missing  unique   Mean   .05   .10   .25   .50   .75   .90   .95
743679      0     15  16.29     0     0    10    10    22    30    40

      0    10    21    22    30    40    50    60    71    72    81    82    91    92    93
Frequency 89727 331302 111495 91177 68870 29819 13597 4399 1341 694 392 476 270 105 15
%         12    45    15    12     9     4     2     1     0     0     0     0     0     0
  
```

```

-----
wp
      n missing  unique   Mean   .05   .10   .25   .50   .75   .90   .95
743679      0    422   4.43     0     0     1     1     3     7    12

lowest :   0    1    2    3    4, highest: 1792 1836 1990 2008 2014
  
```

```

-----
SomVanomzet_excl_btw
      n missing  unique   Mean   .05   .10   .25   .50   .75   .90   .95
729905  13774 129518 564893     0   1835  12493  46757 163882 580545 1342257

lowest : -2576485 -1190338 -500000 -451252 -372191, highest: 801725000 812306000
826774000 867198000 949745025
  
```

```

-----
SomVangeimputeerd_excl_btw
      n missing  unique   Mean   .05   .10   .25   .50   .75
.90      .95
  
```



59063 684616 25525 554787 0.0 452.4 8481.5 45705.0 88602.5
427158.2 1476392.9

lowest : -6995 -6535 -6284 -5973 -214, highest: 259493849 271757181
292643356 293007964 555940337

isWaarneemgroep

	n	missing	unique	Mean
	743679	0	1	0

bron

	n	missing	unique
	743679	0	2

Register (735707, 99%), Response (7972, 1%)



Quality Report Card for Administrative data



Version 0.3

BLUE-ETS project

Data source studied

Insurance Policy record Administration (sometimes also referred to as Social Security Database)
--

Document Version

Version	Adaptations	Responsible	Date
0.2	Based on the initial idea and the proposal included in BLUE-ETS WP4 deliverable 2 a Quality Report Card for Administrative data is proposed	Statistics Netherlands	22/06/2012
0.3	Adjusted version as result of WP8 case studies	Statistics Netherlands	18/01/2013

Document description

This document includes a Quality report Card for Administrative data used for the evaluation of the statistical usefulness of administrative data. The list focuses on the essential quality dimensions and indicators identified in deliverables 1 and 2 of BLUE-ETS workpackage 4. Evaluation is performed by a user at a National Statistical Institute (NSI). The outcome of the evaluation assists the user in the decision to use the data for statistics production.

Instructions

To report the evaluation findings of the quality indicators identified for administrative data a quality report card has been developed. It is named a Quality Report Card for Administrative data (QRCA) and is used to standardize reporting of the evaluation results obtained. Both the findings of automated (scripted) and non-automated quality indicators methods can be noted in the QRCA.

The users starts filling in the QRCA by reporting the results for the Technical Checks dimension, followed by the Time-related, Completeness, Accuracy and Integrability dimension findings. For each dimension, the results of all the measurements methods that can be performed have to be noted down. Quantitative scores can be noted directly. Qualitative score need to be expressed by the signs +, o, - and ? which are used to identify good, reasonable, poor and unclear. Intermediary scores are created by combing symbols with a slash (/) as a separator. This is identical the scores used in the metadata-checklist of Daas and Ossen (2011). Additional space is included to write down remarks.

For each dimension, the evaluation ends with an overall conclusion. This needs to be filled-in by the user. Possible dimensional findings are ‘No problems found’ (green checkbox), ‘Some minor issues observed’ (orange checkbox) and ‘Serious problems detected’(red checkbox). When serious problems are found, this should be noted and –very likely- evaluation must stop. Any instructions included should be followed. When minor issues are found, this also need to be noted but evaluation can continue. When no problems are found evaluation should also continue of course.

In case of serious issues or when all dimensions have been evaluated, the general findings section should be filled in. First, the scores found at the dimensional level are copied and (if needed) converted to the signs +, o, - and ? (see above). If needed, the user can provide additional information at the object or variable level for each dimension. Here also, additional space is included to write down remarks. The summary ends with an overall conclusion for the quality of the data studied which has to be filled in by the user. Plots or other graphical representations of findings can be added to the document if needed.

Information of the NSI-representative who fills in the report

Full name	Piet Daas, Martijn Tennekes,
Position	Methodologists
Department	Methodology and Quality
Phone number	+31-not shown for privacy reasons--
E-mail address	<not_shown>@cbs.nl
Date on which the report card was completed	21 December 2012

1. Technical checks

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Score	Remarks
1.1	Readability		+	Data stored in a SQL-database was used
1.2	File declaration compliance		+	All correct
1.3	Convertability		+	The SQL-data was exported and stored in ff-format prior to analysis. The latter format enables fast access of large data files in R (see ff-package info).

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
Data needed to be converted to ff-format to enable the evaluation of around 20 million records in the Social Security Register by the R-scripts.	No problems found	<input checked="" type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here: A tableplot was created to additional check the dataset studied (figure 1). The visualise (figure 2) and describe (table 1) scripts were also used to provide an overview of the data. See appendix.		

- Only continue when the GREEN or ORANGE marked area is checked.
 - When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted and the section or person responsible for receipt of the data needs to be contacted.

2. Time-related dimension

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Measurement method used	Score	Remarks
2.1	Timeliness		Contacted expert	+	Info not in dataset. Script could not be applied.
2.2	Punctuality		Contacted expert	+	Info not in dataset. Script could not be applied
2.3	Overall time lag		Contacted expert	+	Info not in dataset. Script could not be applied
2.4	Delay		Contacted expert	o	Info not in dataset. Script could not be applied
2.5	Dynamics of objects	Objects	birth, death1, death2, change, change_fast (R-scripts)	b: 1.8% d1: 1.1% d2: 1.1% c: 0.8% cf: 0.8%	Compared unique identifiers <i>February 2010</i> with <i>March 2010</i>
2.6	Stability of variables	Variables	changed_value, unchanged_value (R-scripts)	c: 59% u: 41%	Compared wages in <i>February 2010</i> with <i>March 2010</i>

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
Some scripts could not be tested on the dataset as the dates (of submission and receipt) were not included.	No problems found	<input checked="" type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here: Additional information on the time-related issues were obtained from the Statistics Netherlands expert for the dataset studied. When comparing objects the anonymized variable 'RINPersoon' was used.		

- Only continue when the GREEN or ORANGE marked area is checked.
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.

3. Completeness

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Measurement method used	Score	Remarks
3.1	Undercoverage	Objects	undercoverage (R-script)	u: 1.7%	Compared unique identifiers <i>February 2010</i> with <i>March 2010</i>
3.2	Overcoverage	Objects	overcoverage (R-script)	o: 2.8%	Compared unique identifiers <i>February 2010</i> with <i>March 2010</i>
3.3	Selectivity	Objects	tabplot (R-package)	+	See tableplot (figure 1)
3.4	Redundancy	Objects	redundancy (R-script)	r_a: 5.2% r_b: 95%	Compared unique identifiers for persons (a) and companies (b) for <i>February 2010</i>
3.5	Missing values	Variables	redundancy (R-script)	r: 0.004%	Compared age for <i>February 2010</i>
3.6	Imputed values	Variables		?	Info not in dataset. Expert contacted.

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
No serious problems were found but the imputation indicator (3.6) needs to be answered to assure this.	No problems found	<input type="checkbox"/>
	Some minor issues observed	<input checked="" type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here: The contact person for the data source at UWV needs to be contacted to obtain additional info on the potential use of imputed values.		

- Only continue when the GREEN or ORANGE marked area is checked.
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.

4. Accuracy

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Measurement method used	Score	Remarks
4.1	Authenticity	Objects	checkFormat	c: 0.7% had F included	For <i>February 2010</i> unique identifiers for companies containing were checked is number started with an F (this is allowed)
4.2	Inconsistent objects	Objects	relation	1 person linked to 563 companies	For <i>February 2010</i> unique identifiers for persons and companies were compared
4.3	Dubious objects	Objects	relation	1 person linked to 52 and 1 to 72 companies	For <i>February 2010</i> unique identifiers for persons and companies were compared
4.4	Measurement error	Variables			Can not be determined
4.5	Inconsistent values	Variables	edit_rules	e:0.004%	Nationality unknown (9999)
4.6	Dubious values	Variables	edit_rules	e: 0.7%	For <i>February 2010</i> checked wages of zero or less

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
Several persons were found to be associated with 52 companies or more. This issue has been reported to the expert for the data source studied.	No problems found	<input type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input checked="" type="checkbox"/>
Write additional remarks here: Evaluation was continued to test the checklist. We could find no way of testing measurement errors for this dataset (without using another source of information)		



- *Only continue when the GREEN or ORANGE marked area is checked.*
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- *When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.*

5. Integrability

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Indicators	Level	Measurement method used	Score	Remarks
5.1	Comparability of objects	Objects	coverage	c: 97%	Compared unique identifier for persons for <i>February 2010</i> and <i>March 2010</i>
5.2	Alignment of objects	Objects	coverage	c: 97%	Compared unique identifier for persons for <i>February 2010</i> and <i>March 2010</i>
5.3	Linking variable	Variables	Redundancy, checkFormat	r: 5.3% c: 100%	Checked unique identifier for persons in <i>February 2010</i>
5.4	Comparability of variables	Variables	changed_value, unchanged_value	c: 59% u: 41%	Compared wages in <i>February 2010</i> with <i>March 2010</i>

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
Source needs to be compared with (other) reference data.	No problems found	<input type="checkbox"/>
	Some minor issues observed	<input checked="" type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here: Evaluation needs to be repeated by comparing objects and variables from the data source studied with those in another (reference) source. Because this was out of scope of the project, scripts were run on different reporting periods.		

- Only continue when the GREEN or ORANGE marked area is checked.
 - When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.

General findings

Data source studied

Insurance Policy record Administration (sometimes also referred to as Social Security Database)
--

Dimensional scores

The symbols used to indicate qualitative scores are: good (+), reasonable (o), poor (-) and unclear (?); if needed, symbols can be combined with a slash (/) as a separator.

	Data dimensions	Level	Score	Remarks
1	Technical Checks	Overall	+	
2	Time related	Overall	o	Expert info was used
		Objects	+	
		Variables	+/o	
3	Completeness	Overall	+ (?)	Occurrence of imputed values unknown
		Objects	+	
		Variables	+ (?)	Because of imputed value issue
4	Accuracy	Overall	-	Inconsistent object detected
		Objects	-	Persons working for 52 or more companies detected
		Variables	+ (?)	Measurement error not determined
5	Integrability	Overall	o	Info from another data source was needed
		Objects	o	Info from another data source was needed
		Variables	o	Info from another data source was needed

Overall conclusion	Overall score	
<i>Write additional remarks here:</i> Dubious object issues needs to be resolved	Negative	<input checked="" type="checkbox"/>
	Neutral	<input type="checkbox"/>
	Positive	<input type="checkbox"/>

Appendix

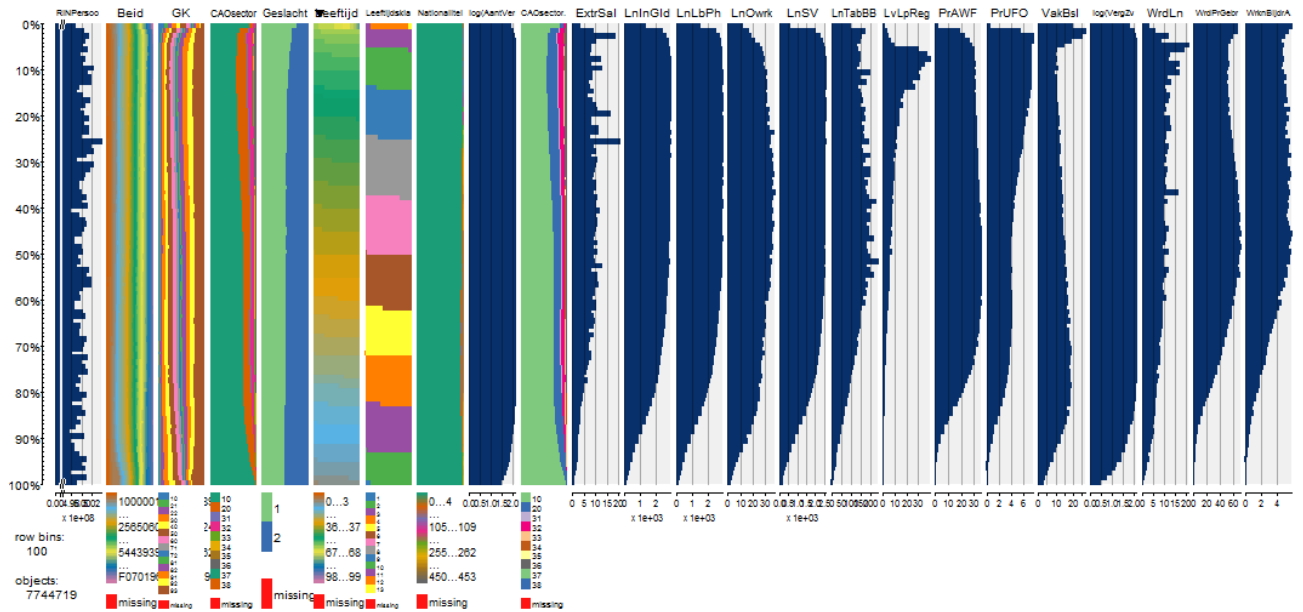


Figure 1: Tableplot for the period February 2010 for the data source studied. Data is sorted on Age (“Leeftijd”). More detailed information on the findings revealed by tableplots are described in the NTTs-paper “On the exploration of high cardinality categorical data” of Tennekes and de Jonge (2013).

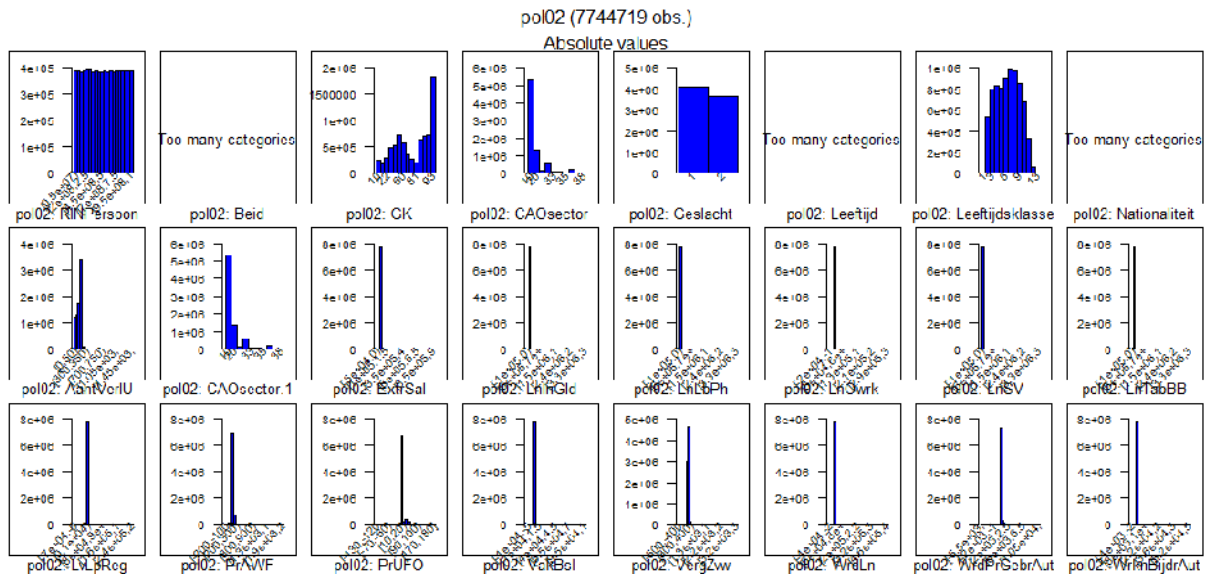


Figure 2: Result of the visualize script for the period February 2010 for the data source studied.

Table 1. Output of the describe script for the period February 2010 for the data source studied.

```

24 Variables          7744719 Observations
-----
RINPersoon
      n  missing  unique  Mean      .05      .10      .25      .50      .75
.90   .95
      7744719          0   7336385    5e+08  50032250  99920165  249784487  500012277  750210819
900006815 950119983

lowest : 0.00e+00 2.00e+00 1.55e+02 2.37e+02 3.89e+02, highest: 1.00e+09 1.00e+09 1.00e+09
1.00e+09 1.00e+09
-----
Beid
      n  missing  unique
7744719      0  422731

lowest : 10000011 10000038 10000046 10000062 10000089, highest: F1047515 F1047517 F1047518
F1047519 F1047520
-----
GK
      n  missing  unique
7744719      0      14

      10      21      22      30      40      50      60      71      72      81      82      91
92      93
Frequency 248924 173018 279780 484760 528073 733173 581968 360236 261226 213224 631509 706052
714806 1827970
%          3      2      4      6      7      9      8      5      3      3      8      9
9      24
-----
CAOsector
      n  missing  unique
7744719      0      10

      10      20      31      32      33      34      35      36      37      38
Frequency 5332384 1357837 129352 546018 67010 64243 5651 215167 14389 12668
%          69      18      2      7      1      1      0      3      0      0
-----
Geslacht
      n  missing  unique
7744719      0      2

```

1 (4089836, 53%), 2 (3654883, 47%)

Leeftijd

	n missing	unique
7744690	29	97

lowest : 0 3 7 8 9 , highest: 97 98 99 100 109

Leeftijdsklasse

	n missing	unique
7744719	0	13

	1	2	3	4	5	6	7	8	9	10	11	12
13												
Frequency	8661	535361	796755	828805	803359	901202	983890	971397	849282	672139	324579	63119
6170												
%	0	7	10	11	10	12	13	13	11	9	4	1
0												

Nationaliteit

	n missing	unique
	0	224

lowest : 0 1 2 3 4 , highest: 454 455 499 500 9999

AantVerlU

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	5786	119.9	18	32	80	140	160	173	174

lowest : 0.00 0.07 0.08 0.10 0.14, highest: 923.59 947.00 968.00 1137.00 1460.00

CAOsector.1

	n missing	unique
7744719	0	10

	10	20	31	32	33	34	35	36	37	38
Frequency	5332384	1357837	129352	546018	67010	64243	5651	215167	14389	12668
%	69	18	2	7	1	1	0	3	0	0

ExtrSal

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	45129	7.648	0	0	0	0	0	0	0

lowest : -3410 -2815 -2637 -2500 -2288, highest: 256561 290000 307721 505649 900000

LnInGld

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	742580	2311	126.0	258.2	974.3	2004.4	3031.8	4320.2	5455.0

lowest : -43858 -14335 -7567 -6735 -5456, highest: 2463476 2494734 2657000 3030711 3318150

LnLbPh

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	803234	2362	142.8	283.1	1011.3	2039.5	3086.1	4389.9	5572.0

lowest : -2093.6 -1518.0 -1268.2 -1243.6 -633.7, highest: 2465266.0 2494977.0 2657000.0 3032631.2 3314413.5

LnOwrk

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	124901	27.73	0.0	0.0	0.0	0.0	0.0	0.0	106.3

lowest : -14303 -13566 -13302 -11921 -6300, highest: 48585 67773 68250 71300 296000

LnSV

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	732730	2092	63.95	180.22	828.56	1832.29	2778.74	3919.54	4940.87

lowest : -12020 -11485 -10027 -9722 -6593, highest: 2463524 2494782 2657000 3030673 3312744

LnTabBB

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	246936	136.6	0.0	0.0	0.0	0.0	0.0	133.8	485.7

lowest : -69191 -57566 -50890 -43235 -40892, highest: 2448677 2479927 2657000 2999963 3304500

LvLpReg

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	46675	7.382	0	0	0	0	0	0	0

lowest : -69191 -44431 -36831 -32489 -30210, highest: 100000 102483 147110 158771 250000

PrAWF

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	23284	26.22	0.00	0.00	0.00	0.01	44.63	93.49	112.04

lowest : -167.0 -121.9 -118.8 -116.0 -112.0, highest: 900.4 1173.0 1666.9 1806.4 1944.6

PrUFO

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	7114	4.101	0.00	0.00	0.00	0.00	0.00	22.33	33.52

lowest : -128.46 -127.84 -126.81 -124.70 -68.03, highest: 131.91 143.49 146.27 175.42 177.25

VakBsl

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	82228	13.88	0.00	0.00	0.00	0.00	0.00	0.00	6.05

lowest : -8792 -7827 -4000 -2671 -2487, highest: 23433 25000 26425 80061 96864

VergZvw

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	29470	117.2	0.00	0.00	57.49	128.60	190.77	194.99	194.99

lowest : -491.9 -441.0 -410.3 -306.1 -203.4, highest: 1103.7 1110.6 1176.0 1511.3 3264.1

WrdLn

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	55161	8.905	0	0	0	0	0	0	0

lowest : -35281 -29995 -27520 -24846 -18921, highest: 363333 386641 419323 448931 464785

WrdPrGebrAut

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	102345	45.25	0	0	0	0	0	0	487

lowest : -5101 -4218 -2718 -2632 -2411, highest: 6985 7141 8574 10664 10839

WrknBijdrAut

	n missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
7744719	0	46254	3.795	0	0	0	0	0	0	0

lowest : -3983 -3770 -3707 -3665 -3542, highest: 6985 10664 11530 15000 52370

Appendix B: Evaluation results of Italian Institute of National Statistics



Funded under Socio-economic Sciences & Humanities



BLUE- Enterprise and Trade Statistics

BLUE-ETS

SP1-Cooperation-Collaborative Project
Small or medium-scale focused research project
FP7-SSH-2009-A
Grant Agreement Number 244767
SSH-CT-2010-244767

Deliverable 8.2 – Istat (Italy)

Title: Methodological case study for testing and evaluating WP4 input data quality indicators - Measuring quality of the Italian social security data as input for the statistical production process

Authors: Grazia Di Bella, Lorena Galiè

With contributions of Daniela Bonardo, Fulvia Cerroni, Valentina Talucci

28 December 2012

Istat Report for WP8 Deliverable

**Methodological case study for testing and evaluating WP4 input data quality indicators:
Measuring quality of the Italian social security data as input for the statistical production
process**

Summary

In this report the Italian methodological case study is described. In Section 1 the WP4 activity will be briefly traced, in Section 2 a description of the administrative data source used for the test is reported. The working method adopted to compute indicators is presented in Section 3. Results are presented in Section 4 reporting evidences aimed at highlighting the indicators robustness. Finally some conclusions are given.

INTRODUCTION.....	46
1. BACKGROUND.....	47
2. DESCRIPTION OF THE SOURCES USED FOR THE ISTAT CASE STUDY	59
3. WORKING METHOD	62
4. CASE STUDY RESULTS.....	63
4.1. TECHNICAL CHECKS	63
4.2. INTEGRABILITY.....	63
4.2.1. Integrability of objects: Comparability of objects.....	63
4.2.1.1. Comparability of objects – Method 1	65
4.2.1.2. Comparability of objects – Method 2	67
4.2.2. Integrability of objects: Alignment of Objects.....	70
4.2.2.1. Alignment of objects - Method 1.....	71
4.2.2.2. Alignment of objects - Method 2.....	74
4.2.3. Integrability of variables – Linking variables	75
4.2.3.1. Linking variables - Method 1	76
4.2.3.2. Linking variables - Method 2	77
4.2.3.3. Linking variables - Method 3	79
4.2.4. Integrability of variables – Comparability of variables.....	81
4.3. ACCURACY	82
4.3.1. Accuracy of objects: Authenticity.....	82
4.3.2. Accuracy of objects: Inconsistent objects	83
4.3.3. Accuracy of objects: Dubious objects.....	84
4.3.3.1. Dubious objects – Method 1	85
4.3.4. Accuracy of variables: Measurement error	86
4.3.5. Accuracy of variables: Inconsistent values	87
4.3.5.1. Inconsistent values – Method 1	87
4.3.6. Accuracy of variables: Dubious values.....	89
4.3.6.1. Dubious values – Method 1	89
4.4. COMPLETENESS	91
4.4.1. Completeness of objects: Undercoverage	91
4.4.1.1. Undercoverage – Method 1	92
4.4.2. Completeness of objects: Overcoverage	93
4.4.2.1. Overcoverage.....	94
4.4.3. Completeness of objects: Selectivity	95
4.4.4. Completeness of objects: Redundancy.....	96
4.4.4.1. Redundancy – Method 1.....	97

4.4.4.2.	Redundancy – Method 2.....	99
4.4.4.3.	Redundancy – Method 3.....	101
4.4.5.	Completeness of variables: Missing values	102
4.4.5.1.	Missing values - Method 1	102
4.4.5.2.	Missing values - Method 2	105
4.4.5.3.	Missing values - Method 3	106
4.4.6.	Completeness of variables: Imputed values.....	107
4.5.	TIME RELATED	108
4.5.1.	Dynamics of objects.....	108
4.5.1.1.	Dynamics of objects – Method 1	109
4.5.1.2.	Dynamics of objects – Method 2.....	111
4.5.1.3.	Dynamics of objects – Method 3.....	112
4.5.2.	Stability of variables	113
4.5.2.1.	Stability of variables – Method 1.....	114
4.5.2.2.	Stability of variables – Method 2.....	116
4.5.2.3.	Stability of variables – Method 3.....	117
	CONCLUSIONS	118
	REFERENCES.....	118

Introduction

Administrative data were added in the last few years as a further source, next to the data collected from sample survey and census, for the production of official statistics in NSIs.

This synergy allows to improve the statistical production process and reduce the so-called "statistical burden" among economic agents.

Many statistical processes in EU, especially for producing business statistics, use different types of administrative sources. An interesting overview on which kind of Administrative Data are mostly used has been conducted by AdminData ESSnet - WP1 "Overview of existing practices in the use of administrative data for producing business statistics over Europe" (Costanzo et al., 2011)¹.

Some administrative sources entered the production process in a consolidated manner but they must be constantly monitored for two main reasons: a) their statistical use is secondary and regulatory changes can produce significant breaks in the periodical deliveries and may impact the statistics production process; b) before the data are introduced in the production process, a check procedure must be performed to make sure that there are no unexpected errors.

Other administrative sources are currently being evaluated to verify their usability in the production process.

In order to support these actions, it is necessary to define new tools. These tools must ensure the timeliness and standardize as far as possible, the evaluation of sources.

This is the WP4 contribution: to develop data input quality indicators for monitoring administrative data quality entering the production process or for evaluating their possible usability. The main goal is to define a new comprehensive quality-indicator instrument, a Quality Report Card (QRC) for Administrative data that can be generally used by NSIs.

In the first WP4 Deliverable (Daas et al., 2011a) a first indicators proposal has been made within the relevant quality dimensions, in the second WP4 Deliverable (Daas et al., 2011b) a shared quality framework and a final list of indicators has been presented.

The last step of WP4 activity is testing quality dimensions and corresponding quality indicators on a specific administrative data to verify their robustness.

As far as Istat is concerned, the measurement methods of the quality indicators have been tested on the Italian Social Security Data, an important administrative source the use of which has recently been enhanced in Istat for the statistical production process.

¹ The ESSnet AdminData on *The Use Of Administrative And Accounts Data For Business Statistics* - under the MEETS programme - aims to find common ways for use of admin data for business statistics. More information on the web site: <http://essnet.admindata.eu/>.

1. Background

In the development of quality indicators associated with the use of administrative data for statistical purposes, it is possible to distinguish three types of indicators:

1. **Input quality indicators:** to define the input quality of the statistical production process
2. **Process quality indicators:** to measure the quality of the production process that uses administrative data to produce statistics
3. **Output quality indicators:** to measure the output quality of the business statistics involving administrative data, taking input and process into account.

WP4 activity focused on developing type 1 indicators on Input quality.

It is important to remember that the project ESSnet AdminData - WP6 “Development of quality indicators” is centred on indicators of type 3 (Frost et al., 2010). Since the beginning, a coordination between the two working groups has been established to better reach their goals².

For the input quality indicators, following the Statistics Netherlands quality framework (Daas et al., 2009), it is possible to face the Quality evaluation of Admin Data Sources with a hierarchical approach. Two levels are considered:

- a **Metadata level** defined through the dimensions of Relevance, Privacy and security (legal basis), Clarity, Comparability, Unique keys,...
- a **Data level** related to the data quality (facts).

WP4 investigated quality dimensions and indicators for the **Data level**.

Another useful specification of Input quality indicators derived within Blue Ets WP4 work (Daas et al., 2011) is about the evaluation of the quality of an Administrative source. It can vary depending on the value of the additional information brought to the specific statistical production process.

A general quality assessment not considering the specific additional information to a statistical process is referred to as **Data Source Quality (DSQ)**, otherwise it is called **Input Output oriented Quality (IOQ)**. Where possible, the performed indicators are also classified under this criterion.

One last remark: obviously the indicators presented are producer statistics-oriented and not user statistics-oriented since they have the aim to support the statistical production within NSIs.

Five Quality components were selected to determine the input quality of AdminData sources.

² For more information on ESSnet AdminData WP6 results, see the website <http://essnet.admindata.eu/>.

Table 1.1 - Quality dimensions and indicators

DIMENSION	LEVEL	INDICATORS
TECHNICAL CHECKS	-	Technical usability of the file and data in the file
INTEGRABILITY	Objects/Variables	Extent to which the data source is capable of undergoing integration or of being integrated.
ACCURACY	Objects/Variables	The extent to which data are correct, reliable and certified
COMPLETENESS	Objects/Variables	Degree to which a data source includes data describing the corresponding set of real-world objects and variables
TIME-RELATED DIMENSION	Objects/Variables	Indicators that are time and/or stability related

Within each of the five dimensions an additional information concerns indicators specific for objects (generally units and events) and for variables. This distinction is very useful for the exploitation of administrative sources. In fact, the units provide an objective assessment as they are not defined and planned a priori but derived a posteriori for statistical purposes. So, for each dimension, with the exception of the Technical Checks one, a level is defined considering indicators for objects and for variables.

In the following tables quality indicators defined in the second WP4 Blue Ets Deliverable (Daas et al., 2001b) are shown. Each indicator is classified with respect to its focus, between Data Source Quality (DSQ) and Input Output oriented Quality (IOQ).

Measurement methods selected and proposed by WP4, are also reported. Only those tested for the Istat case study are marked (T = tested).

Table 1.2 TECHNICAL CHECKS - Indicators

	Level	Indicators	Description	Focus
TECHNICAL CHECKS	-	Readability	Accessibility of the file and data in the file	DSQ / IOQ
	-	File declaration compliance	Compliance of the data in the file to the metadata agreements	DSQ/ IOQ
	-	Convertibility	Conversion of the file into the NSI-standard format	DSQ/ IOQ

Table 1.3 TECHNICAL CHECKS - Measurement methods

Dimension indicators	Level	Measurement methods	T
Readability	-	% of deliveries (or files) of the total deliveries with an unknown extension, that are corrupted, or cannot be opened	
		% of the total file which is unreadable (in size (MB/GB) or number of readable file records)	
File declaration compliance	-	% of variables in the current delivery that differ from the metadata lay-out agreed upon in: i) formats and names ii) variable and attribute content iii) categories defined for categorical variables iv) ranges for numerical variables (if applicable, e.g. for age: 0-120)	
Convertibility	-	% of objects with decoding errors or corrupted data	

Table 1.4 INTEGRABILITY DIMENSION - Indicators

	Level	Indicators	Description	Focus
INTEGRABILITY	Objects	Comparability of objects	Similarity of objects in source with the objects used by NSI	IOQ
	Objects	Alignment of objects	Linking-ability (align-ability) of objects in source with those of NSI	IOQ
	Variables	Linking variable	Usefulness of linking variables (keys) in source	IOQ
	Variables	Comparability of variables	Proximity (closeness) of variables	IOQ

Table 1.5 INTEGRABILITY DIMENSION - Measurement methods

Dimension indicators	Level	Measurement methods	T
Comparability of objects	Objects	% of identical objects = (Number of objects with exactly the same unit of analysis and same concept definition as those used by NSI) / (Total number of relevant objects in source) x 100	✓
		% of corresponding objects = (Number of objects that, after harmonization, would correspond to the unit needed by NSI) / (Total number of relevant objects in source) x 100	✓
		% of incomparable objects = (Number of objects that, even after harmonization, will not be comparable to one of the units needed by NSI) / (Total number of relevant objects in source) x 100	
		% of non-corresponding aggregated objects = (Fraction of objects of interest at an aggregated level in source 1 - fraction of objects of interest at the same aggregated level in source 2) x 100	
Alignment of objects	Objects	% of identical aligned objects = (Number of objects in the business register with exactly the same unit of analysis and same concept definition as those in the source) / (Total number of relevant objects in business register) x 100	✓
		% of corresponding aligned objects = (Number of objects in the business registers that, after harmonization, correspond to units or parts of units in the source) / (Total number of relevant objects in business register) x 100	✓

Table 1.5 INTEGRABILITY DIMENSION - Measurement methods (continued)

Dimension indicators	Level	Measurement methods	T
Alignment of objects	Objects	% of non-aligned objects = (Number of objects in the business register that, even after harmonization of the objects in the source, can not be aligned to one of the units in the source) / (Total number of relevant objects in business register) x 100	
		% of non-aligned aggregated objects = (Fraction of objects of interest at an aggregated level in source 1 that can not be aligned + fraction of objects of interest at the same aggregated level in source 2 that can not be aligned) x 100	
Linking variable	Variables	% of objects with no linking variable = (Number of objects in source without a linking variable) / (Total number of objects in the source) x 100	✓
		% of objects with (a) linking variable(s) different from the one(s) used by NSI = (Number of objects in source with (a) linking(s) variable different from the one used by the NSI) / (Total number of objects with (a) linking variable(s) in the source) x 100	✓
		% of objects with correctly convertible linking variable(s) = (Number of objects in the source for which the original linking variable can be converted to one used by the NSI) / (Total number of objects with a linking variable in the source) x 100	✓
Comparability of variables	Variables	Use statistical data inspection methods to compare the totals of groupings of specific objects for variables in both sources. Graphical methods that can be used are a bar plot and a scatter plot. Distributions of values can also be compared.	
		The Mean Absolute Percentage Error (MAPE) that measures the mean of the absolute percentage error. MAPE has a lower bound of zero but has no upper bound. Alternatively the symmetric MAPE could be used. This method measures the symmetric mean of the absolute percentage error were the deviation between the percentage distributions is divided by the half-sum of the deviations.	
		A method derived from the chi-square test that evaluates the distributions of the numeric values in both data sets. For categorical data Cramers V could be used.	
		% of objects with identical variable values = (Number of objects in source 1 and 2 with exactly the same value for the variable under study) / (Total number of relevant objects in both sources) x 100	

Table 1.6 ACCURACY DIMENSION - Indicators

	Level	Indicators	Description	Focus
ACCURACY	Objects	Authenticity	Legitimacy of objects	DSQ / IOQ
	Objects	Inconsistent objects	Extent of erroneous objects	DSQ / IOQ
	Objects	Dubious objects	Presence of untrustworthy objects	DSQ / IOQ
	Variables	Measurement error	Correctness of a value with respect to the measurement process	DSQ / IOQ
	Variables	Inconsistent values	Extent of inconsistent combinations of variables values	DSQ / IOQ
	Variables	Dubious values	Presence of implausible values or combinations of values for variables	DSQ / IOQ

Table 1.7 ACCURACY DIMENSION - Measurement methods

Dimension indicators	Level	Measurement methods	T
Authenticity	Objects	% of objects with a non-syntactically correct identification key	✓
		% of objects for which the data source contains information contradictory to information in a reference list for those objects (master list and target list)	
		Contact the data source holder for their % of non-authentic objects in the source	
Inconsistent objects	Objects	% of objects involved in non-logical relations with other (aggregates of) objects	
Dubious objects	Objects	% of objects involved in implausible but not necessarily incorrect relations with other (aggregates of) objects	✓
Measurement error	Variables	% of unmarked values in the data source for each variable (when values not containing measurement errors are marked by administrative data holder)	
		Contact the data source holder and ask the following data quality management questions: - Do they apply any design to the data collection process (if possible)? - Do they use a process for checking values during the reporting phase? - Do they use a benchmark for some variables? - Do they use a checking process for data entry? - Do they use any checks for correcting data during the processing or data maintenance?	

Table 1.7 ACCURACY DIMENSION - Measurement methods *(continued)*

Dimension indicators	Level	Measurement methods	T
Inconsistent values	Variables	% of objects of which combinations of values for variables are involved in non-logical relations	✓
Dubious values	Variables	% of objects with combinations of values for variables are involved in implausible but not necessarily incorrect relations	✓

Table 1.8 COMPLETENESS DIMENSION - Indicators

	Level	Indicators	Description	Focus
COMPLETENESS	Objects	Undercoverage	Absence of target objects (missing objects) in the source	DSQ / IOQ
	Objects	Overcoverage	Presence of non-target objects in the source	IOQ
	Objects	Selectivity	Statistical coverage and representativeness of objects	DSQ / IOQ
	Objects	Redundancy	Presence of multiple registrations of objects	DSQ
	Variables	Missing values	Absence of values for (key) variables	DSQ / IOQ
	Variables	Imputed values	Presence of values resulting from imputation actions by data source holder	DSQ / IOQ

Table 1.9 COMPLETENESS DIMENSION - Measurement methods

Dimension indicators	Level	Measurement methods	T
Undercoverage	Objects	% of objects of the reference list missing in the source	✓
Overcoverage	Objects	% of objects in the source not included in the reference population	✓
		% of objects in the source not belonging to the target population of the NSI	✓
Selectivity	Objects	Use statistical data inspection methods, such as histograms, to compare a background variable (or more than one) for the objects in the data source and the reference population	
		Use of more advanced graphical methods, such as tableplots	
		Calculate the R-indicator for the objects in the source	
Redundancy	Objects	% of duplicate objects in the source (with the same identification number)	✓
		% of duplicate objects in the source with the same values for a selection of variables	✓
		% of duplicate objects in the source with the same values for all variables	✓

Table 1.9 COMPLETENESS DIMENSION - Measurement methods *(continued)*

Dimension indicators	Level	Measurement methods	
Missing values	Variables	% of objects with a missing value for a particular variable	✓
		% of objects with all values missing for a selected (limited) number of variables	✓
		Use of graphical methods to inspect for missing values for variables	✓
Imputed values	Variables	% of imputed values per variable in the source	
		Contact the data source holder and request the percentage of imputed values per variable	

Table 1.10 TIME RELATED DIMENSION - Indicators

	Level	Indicators	Description	Focus
TIME RELATED		Timeless	Lapse of time between the end of the reference period and the moment of receipt of the data source	DSQ / IOQ
		Punctuality	Possible time lag between the actual delivery date of the source and the date it should have been delivered	DSQ / IOQ
		Overall time lag	Overall time difference between the end of the reference period in the source and the moment the NSI has concluded that it can definitely be used	DSQ / IOQ
		Delay	Extent of delays in registration	DSQ / IOQ
	Objects	Dynamics of objects	Changes in the population of objects (new and dead objects) over time	IOQ
	Variables	Stability of variables	Changes of variables or values over time	IOQ

Table 1.11 TIME RELATED DIMENSION - Measurement methods

Dimension indicators	Level	Measurement methods	T
Timeless		Time difference (days) = (Date of receipt by NSI) – (Date of the end of the reference period over which the data source reports)	
		Time difference (days) = (Date of receipt by user) – (Date of the end of the reference period over which the data source reports)	
Punctuality		Time difference (days) = (Date of receipt by NSI) – (Date agreed upon; as laid down in the contract)	
Overall time lag		Total time difference (days) = (Predicted date at which the NSI declares that the source can be used) – (Date of the end of the reference period over which the data source reports)	
Delay		Contact the data source holder to provide their information on registration delays	
		Time difference (days) = (Date of capturing the change in the source by the data source holder) – (Date the change occurred in the population)	
Dynamics of objects	Objects	$\% \text{ Births } t = (\text{Births } t / \text{Total objects } t) \times 100 = (\text{Births } t / (\text{Births } t + \text{Alive } t)) \times 100$	✓
		$\% \text{ Deaths } t = (\text{Deaths } t / \text{Total objects } t) \times 100 = (\text{Deaths } t / (\text{Births } t + \text{Alive } t)) \times 100$	✓
		$\% \text{ Deaths } t-1 = (\text{Deaths } t / \text{Total objects } t-1) \times 100 = (\text{Deaths } t / (\text{Alive } t + \text{Deaths } t)) \times 100$	✓
Stability of variables	Variables	Use statistical data inspection methods to compare the values of specific variables for persistent objects in different deliveries of the source. Graphical methods that can be used are a bar plot and a scatter plot.	✓
		$\% \text{ of Changes} = (\text{Number of objects with a changed value} / \text{total number of persistent objects with a value filled in for the variable under study}) \times 100\%$	✓
		A correlation statistical method can be used to determine to which extent values changed in the same direction for different object. For categorical data a method such as Cramers V can be used	✓

2. Description of the sources used for the Istat case study

Italian Social Security Data (SSD) used for Istat methodological case study is produced by Inps (Italian Institute of Social Security) and concerns the monthly contribution declarations of employers for employees.

The choice of this archive for the indicators test derives from several elements: (i) it is a complex and big source including more statistical units connected to each other; (ii) its use can be extended to different types of statistical production process; (iii) Istat is particularly interested in it as it will be used to redefine the business register production and to improve its timeliness.

Until last year, Istat used an annual version of the archive that was provided after 18 months from the end of the reference period. Since 2011 INPS provides a monthly version that is currently undergoing a testing phase. The overall provision of monthly data for 2010 became available after about 11 months (November 2011). Data supply is very big, about 160 million records and 45 variables with, however, a high degree of redundancy. In order to test the quality framework and indicators, measurement methods are computed on May 2010 data. This subset of the entire database contains about 13 million records and the same number of variables. The administrative source is also called Emens.

As in the SSD source more types of units are identifiable and these are connected to each other by specific relations, it is necessary to provide a brief description of them.

SSD includes four kinds of objects. Each record can be defined as the **Worker tax position** that is the set of variables useful to define the amount of social security contributions payable for each employee by the employer. Among these variables there are the employer and employee tax codes too, hence. SSD source is a Leed dataset (Linked employer- employee data). Other variables characterizing the Worker tax position are the Type of employment contract (Fixed term/Permanent), Contractual working time (Full/Part-time), Professional status and Type of contribution. The change of any of these characteristics during the month gives rise to a new record concerning the same contract. Due to this database building rule, the information is redundant.

Other administrative units that can be derived from the Tax position are: the Employee, the Employer and the Workplace (municipality).

In the following table, the Administrative units are described and their key variable is reported.

The units Employee and Employer are identified, respectively, through the tax code of the employee and the tax identification number of the company, concerning the Workplace a code called "Belfiore" is used to identify the Italian municipalities. The latter coding system is different from that used by Istat but a trans-coding table is available.

Table 2.1 – Administrative units in the SSD source

Definition	Identification key
<p>Worker tax position– primary unit</p> <p>The set of characteristics useful to define the amount of social security contributions payable for each employee by the employer</p>	Complex key defined by a set of variables
<p>Employee – derived unit</p> <p>A worker who has had at least one pay contributions to INPS as an employee during the month</p>	Employee Tax Code
<p>Employer – derived unit</p> <p>Employer who have made at least a payment contributions for employees in May of 2010 or Employer who has employed at least one regular worker</p>	Enterprise Tax Code
<p>Workplace - derived unit</p> <p>Place where the work is mainly carried out</p>	Belfiore Municipality Code

Among the SSD administrative units there are hierarchical relationships:

- a) an employer may have more employees;
- b) an employee can have more than one tax position with the same employer;
- c) an employee can have more than one tax position with the more than one employer;
- d) a municipality may not be a workplace;
- e) a municipality can be host for more workers.

The statistical sources used as the reference population are the Italian Business Register and Istat Register of Italian municipalities.

The Italian Business Register (BR), named Asia (Archivio Statistico delle Imprese Attive), is produced by Istat and annually updated in accordance with the Business Register Regulation (EC) n. 177/2008.

At the moment Asia does not include (NACE Rev.2)

- Section A - Agriculture, forestry and fishing
- Section O - Public administration and defence, compulsory social security
- Division 94 Activities of membership organisations
- Section T - Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use

Section U - Activities of extraterritorial organisations and bodies
Public corporations
Non-profit institutions

The list of Italian municipalities is produced by Istat and updated twice a year (June 30 and December 31), on the basis of territorial and administrative changes that occurred in the country according to the Classification of territorial units for statistics (NUTS), adopted at the European level. For the application presented here, the list used is the one updated on 1 January 2011. At this point in time, the official number of Italian municipalities is to 8,094 units.

The Emens archive covers data on the social security system for employees in private companies, which are included in the Asia BR as persons employed. As far as the territorial reference is concerned, SSD includes social security contributions payable by enterprises (legal units) resident in Italy.

In Section 4.2.1 the Integrability of objects in the Emens source is considered and the administrative units are accompanied, where possible, by the corresponding statistical unit (Table 4.1).

The SSD source provides a wealth of information useful to describe the occupation in enterprises. In addition to the main variables previously mentioned for describing Worker Tax position, there are the following variables: the number of paid days, the national collective agreement, the date and the reason for hiring, the date and the reason for termination and so on. The hiring date and the termination date are two events (objects) defined in the monthly data with the day of the month.

3. Working method

To test quality indicators on the SSD source we started by performing a preliminary analysis of the indicators and related measurement methods in order to evaluate their applicability to the data. We encountered two types of restrictions. Sometimes the lack of information given by the administrative data holder on the metadata or the data generation process was a critical element. Other restrictions are due to the absence of the reference list (administrative target population or statistical target population).

As a result, for few methods it was not possible to proceed to the testing phase. In other cases, the measurements have not been calculated as evaluating their robustness was not necessary.

In Tables 1.3, 1.5, 1.7, 1.9, 1.11 the tested measurements methods are marked (T = ✓).

Indicators have been also classified between those (Technical checks and some and some Time related dimension indicators) referring to the entire archive, and those (Integrability, Accuracy, Completeness and some Time related dimension indicators) applicable to selected units or variables in the source. In some cases, to calculate the indicators we proceeded involving colleagues responsible for the administrative data management (contacts with the data holder, acquisition, loading of the data source). For the second group indicators it was necessary to perform some preliminary actions:

- Selection of objects in the source to which it was possible and useful to test indicators
- Identification and acquisition of the statistical reference population to compare and match data (Integrability and Completeness dimensions)
- Selection of variables in the source to which it was possible and useful to test indicators

Whereas in the source the units are repeated on multiple records (by construction), the indicators were calculated by counting the number of records or by counting units as needed. In the presentation of the results the criterion adopted is indicated i.e. whether the objects refer to records or to specific units.

For the implementation of the measuring methods we used the statistical software Sas. Initially we evaluated also the hypothesis of performing the processing in R, however, it was decided to use the Sas software because it is more suitable for processing large amounts of data as the SSD source is.

In the next chapter we will present the results of the test carried out. For each indicator and for each measurement method implemented, some preliminary details are provided and a summary table is presented including: the calculation algorithm of the method, the Sas programme implemented, the results obtained and some comments useful for implementation.

4. Case study results

In this Section results of the case study on the administrative source of the Social Security Data (SSD) Emens are presented. WP4 BlueEts Quality indicators (Daas et al, 2011b) are tested with the aim to evaluate their robustness.

4.1. TECHNICAL CHECKS

Technical Checks indicators were not reported since information was not available or their implementation did not involve critical elements in relation to their robustness.

4.2. INTEGRABILITY

Quality indicators of the Integrability dimension are specifically suitable for the statistical use of administrative data as they measure the extent to which the data source can be integrated into the statistical production process of an NSI.

The Integrability indicators refer to usability of units and variables. Among the latter, the linking variables play a special role because they allow to combine data with other micro data sources.

4.2.1. Integrability of objects: Comparability of objects

Comparability refers to the similarity of the objects in the administrative source with those used by NSI. To test indicators, administrative objects are analyzed with respect to their degree of comparability with statistical objects:

- *identical objects*: objects with exactly the same unit of analysis and same concept definition as those used by NSI
- *corresponding objects*: objects that after harmonization would correspond to the statistical unit
- *incomparable objects*: objects that even after harmonization will not be comparable to one of the units needed by NSI.

In Table 4.1 Administrative units, already described in Section 2, are placed side by side with the similar statistical units.

Table 4.1 –Administrative units in the SSD source and Reference statistical units used in Istat

Administrative Units		Integrability level	Statistical Units	
Definition	Identification key		Definition	Identification key
<p>Worker tax position</p> <p>The set of characteristics useful to define the amount of social security contributions payable to INPS for each employee by the employer</p>	Complex key defined by a set of variables	<p>Corresponding</p> <p>↔</p>	<p>Employment relationship</p> <p>A formal agreement between an enterprise and a person, whereby the person works for the enterprise in return for remuneration</p>	- (Employment relationships register not available)
<p>Employee – derived unit</p> <p>A worker for which there is at least one social security contribution paid to INPS as an employee during the month</p>	Employee Tax Code	<p>Identical</p> <p>↔</p>	<p>Employee in Enterprise</p> <p>Person who works for an Enterprise on the basis of a contract of employment and receives compensation</p>	- (Employees register not available)
<p>Employer – derived unit</p> <p>Employer who has employed at least one regular worker</p>	Enterprise Tax Code	<p>Identical</p> <p>↔</p>	<p>Enterprise</p> <p>Enterprise in Business Register</p> <p>Enterprise with Employees</p> <p>Enterprise with employment >0 in Business Register</p>	Enterprise Tax Code
<p>Workplace - derived unit</p> <p>Place where the work is mainly carried out</p>	Belfiore Municipality Code	<p>Corresponding</p> <p>↔</p>	<p>Work Municipality</p> <p>Italian Municipalities</p>	Istat Municipality Code

For the Employee and Employer units there is a similarity (identical objects) between administrative and statistical concepts and they have in the administrative source the same identification variable used by Istat. The units Work Tax position and Municipality are not directly comparable with statistical units of interest. But, after a treatment process, it is possible to integrate them. In particular, the Work Tax position can be used for the identification of the statistical unit Employment relationship (or contract of employment or job), however, we do not have a statistical register for it. About the administrative unit of the Working place it is possible to use an external table to harmonize administrative and statistical units.

To measure the comparability of objects we tested Method 1: percentage of *identical* objects in administrative source (with exactly the same unit of analysis and same concept definition) and Method 2: percentage of *corresponding* objects in administrative source (with the same unit of analysis and same concept definition, only after harmonization).

4.2.1.1. Comparability of objects – Method 1

The method 1 has been tested on the Enterprise unit by using the Asia BR as reference statistical population that is the total business population active in the year (at least one day). These amounted to 4.525.155 units.

The presence of common identification keys between the two lists (tax code) made the record linkage possible.

It should be noted that errors in key variables can affect the outcome of the indicator, so it is necessary to check the results of the Integration indicator for the linking variable (§ 4.2.3).

Comparability of objects – Method 1

Algorithm

$I_{4.2.1.1} = \text{\% of identical objects} = (\text{Number of objects with exactly the same unit of analysis and same concept definition as those in the statistical population}) / (\text{Number of relevant objects in source}) \times 100$

Object refers to Enterprises

Sas programme

```
/* Comparability of Emens enterprises with Asia enterprises*/
proc sort data=sasrava.emens_2010_maggio out=Emens_nodup nodupkey;
by co_dirt_dichiara;
run;
proc sort data=sasrava.Asia10_attive_definitivo3 out=Asia2010;
by codice_f; /*codice_f is Enterprise Tax Code in Asia Business Register*/
run;
data emens_asia;
merge Emens_nodup (rename=( co_dirt_dichiara=codice_f) in=a ) Asia2010 (in=b);
by codice_f;
if a and b then v=1; /* linked Enterprises*/
if a and not b then v=2; /* not linked Enterprises: in Emens but not in Asia2010 */
if b and not a then v=3; /* not linked Enterprises: in Asia2010 but not in Emens */
run;
proc freq data=emens_asia;
tables v/missing;
run;
```

Results

Number of enterprises in the SSD source with exactly the same unit of analysis and same concept definition as those in the Business Register (Asia all enterprises) = 1.376.387

Number of relevant enterprises in the SSD source = 1.508.321

$I_{4.2.1.1} (\text{Enterprise units, Asia enterprises}) = (1.376.387 / 1.508.321) * 100 = 91,25\%$

Comments for implementation

- To calculate this method it is essential to consider a reference statistical population.
- For the computation, a step of record linkage (join with common key) is requested, errors in the linking variable can affect the measurement result.
- This method is connected to method 2 for the alignment of objects (the numerator is the same) and to that for overcoverage indicator.
- The comparability indicator could be useful to point out time alignment issues

4.2.1.2. Comparability of objects – Method 2

Method 2 has been tested on the Working place (Municipality) unit compared to the official Istat Municipalities Register referred to the 1st of January 2011, as reference statistical population (8.094 units).

The administrative units in Emens are not directly comparable with the statistical units in Istat and they refer to different identification codes (named Belfiore code). But a table is available which make the harmonization possible. Administrative units are involved in (n : 1) with (n ≥ 1) relations with statistical units as reported in the following table.

Relation between administrative units and statistical units (Municipalities)

Belfiore code and statistical code	v. a.	%
1:1	7.760	96,24
2:1	235	2,91
3:1	44	0,55
4:1	14	0,17
5:1	2	0,02
7:1	1	0,01
8:1	2	0,02
9:1	2	0,02
10:1	1	0,01
12:1	1	0,01
15:1	1	0,01
Total	8.063	100,00

In the administrative source there are 8.725 distinct Belfiore codes. Among these, 8.506 have been converted in Istat codes ($8.506/8.725 = 97,49\%$) and corresponding to 8.063 Istat Municipalities. So there are 219 Belfiore codes (2,5%) not convertible.

It is possible to measure the share of codes convertible in NSI's code (in SSD case 97,49%) and the share of codes convertible which correspond to statistical units of the same reference period (100% in our case).

As explained in § 4.2.3.1 (Linking variable quality), in SSD source there are also 7.595 records on 13.375.903 (0,06%) in which Belfiore code is missing (Null).

Comparability of objects – Method 2

Algorithm

$I_{4.2.1.2} = \% \text{ of corresponding objects} = (\text{Number of objects in the source that, after harmonization, present the same unit of analysis and the same concept definition as those in the statistical population}) / (\text{Number of objects in source}) \times 100$

Objects refer to Municipalities

Sas programme

```
/*Conversion programme (not reported)*/
/*Programme to verify the relation between administrative Municipalities and statistical Municipalities */
/* Elimination of records having the same pair of codes Belfiore code – Istat code*/
proc sort data= sasrava.emens_2010_maggio out=Emens_mun_nodup nodupkey;
by co_dirt_codcomune cod_mun; /* cod_mun is the Istat Municipality code*/
where co_dirt_codcomune ne '' and cod_mun not in (''999999');
run;
/*Aggregation of records by Istat Code */
proc summary data=Emens_mun_nodup nway;
class cod_mun;
var co_dirt_anno;
output out=Emens_mun_Istat N=;
run;
/* Number of Belfiore codes per each Istat code */
proc freq data= Emens_mun_Istat;
tables _freq_/missing;
run;
/* Comparability of Emens Municipality (after harmonization with Istat units) */
proc sort data=Comuni_istat_1gen2011;
by codice_comune; /* codice_comune is Istat Code in Municipality Register */
run;
proc sort data= Emens_mun_nodup;
by cod_mun;
run;
data emens_register_mun;
merge Emens_mun_nodup (in=a) Comuni_istat_1gen2011 (rename=( codice_comune=cod_mun) in=b );
by cod_mun;
if a and b then v=1; /* linked Municipalities*/
if a and not b then v=2; /* not linked Municipalities: in Emens but not in Istat Register */
if b and not a then v=3; /* not linked Municipalities: in Istat Register but not in Emens */
run;
proc freq data= emens_register_mun;
tables v/missing;
run;
```

Results

Number of Belfiore codes in the SSD source convertible to Istat codes (that, after harmonization, have the same unit of analysis and the same concept definition as those in the NSI's list) = 8.506

Number of Belfiore codes in the SSD source = 8.725

I_{4.2.1.2} (Municipality unit) = $(8.506 / 8.725) * 100 = 97,49\%$

Comments for implementation

- To calculate this method it is essential to have a reference statistical population.
- For the computation, a procedure of harmonization and a step of record linkage (join with common key) are requested
- This method is connected to method 2 for alignment of objects (the numerator is the same)
- The comparability indicator could be useful to point out time alignment issues

4.2.2. Integrability of objects: Alignment of Objects

The alignment of the objects measures the degree of matching of objects in the statistical population to those of the source.

As for the three Comparability methods, the number of objects *identical*, *corresponding* and *incomparable* are considered. In this case the percentage is computed with respect to the statistical population.

We tested method 1 on the Enterprise unit and method 2 on the Municipality unit.

4.2.2.1. Alignment of objects - Method 1

Method 1 has been tested on the Enterprise unit and, as in the Comparability indicator, we used the Asia BR as reference statistical population. To better define the enterprises population closest to the administrative one, we also considered Asia enterprises with employment >0 that should correspond to “Employers who have paid at least one social security contribution to INPS for employees in May of 2010” (Administrative unit definition in Table 2.1).

On the total number of enterprises in the Asia Register (4.525.155), 1.376.387 are also in the SSD source. Considering only the Asia subpopulation of enterprises with employment >0, we found that on 1.563.129 units, 1.370.863 are also in the SSD sources.

It has to be noted that the number of common units is lesser in the second case, because of some possible time misalignment.

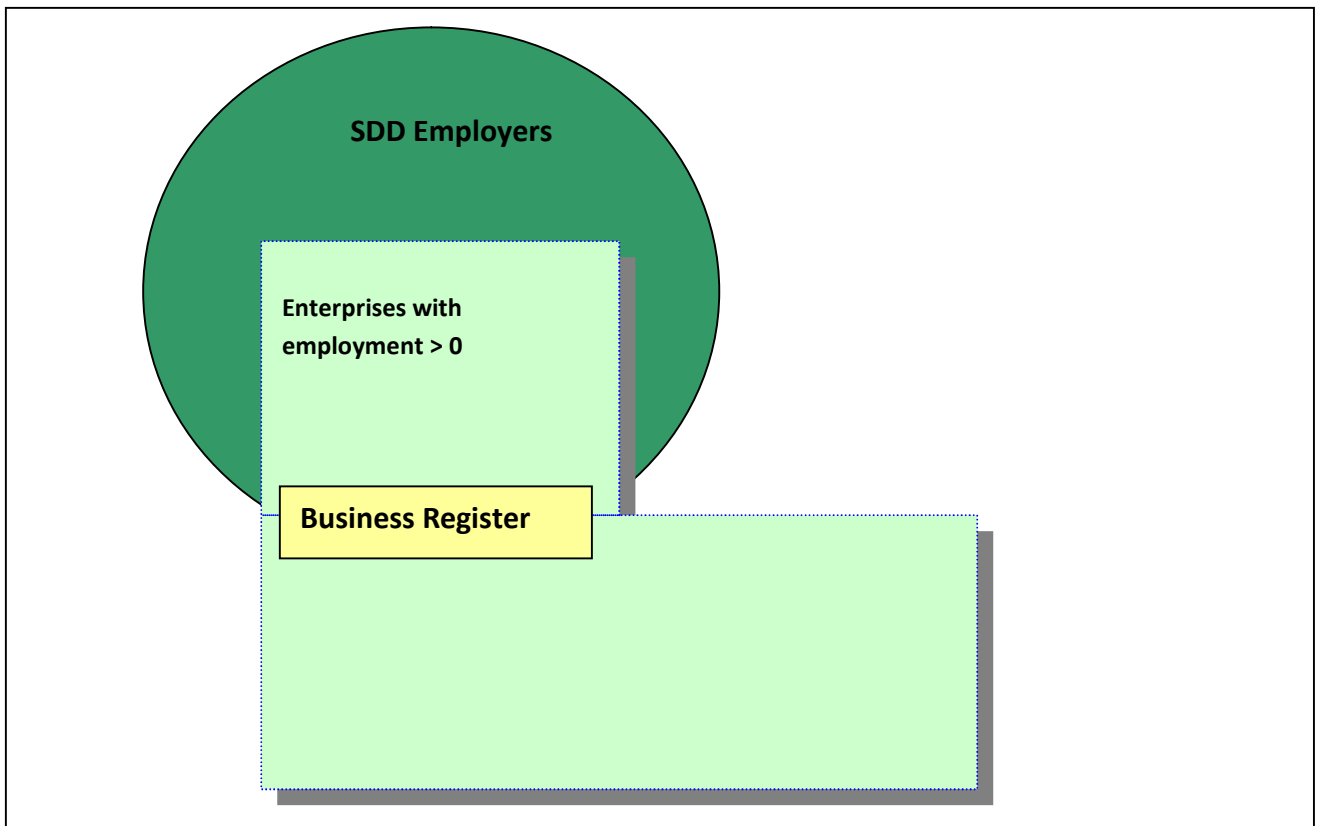


Figure 4.1 – Integrating SSD and Business Register

Alignment of objects - Method 1

Algorithm

$I_{4.2.2.1} = \% \text{ of identical aligned objects} = (\text{Number of objects in the statistical population with exactly the same unit of analysis and the same concept definition as those in the source}) / (\text{Number of relevant objects in the statistical population}) \times 100$

Objects refer to enterprises

Sas programme

```
/* Alignment of Emens enterprises with Asia all enterprises*/
proc sort data=sasrava.emens_2010_maggio out=Emens_nodup nodupkey;
by co_dirt_dichiara;
run;
proc sort data=sasrava.Asia10_attive_definitivo3 out=Asia2010;
by codice_f; /*codice_f is Enterprise Tax Code in Asia Business Register*/
run;
data emens_asia;
merge Emens_nodup (rename=( co_dirt_dichiara=codice_f) in=a ) Asia2010 (in=b);
by codice_f;
if a and b then v=1; /* linked Enterprises*/
if a and not b then v=2; /* not linked Enterprises: in Emens but not in Asia2010 */
if b and not a then v=3; /* not linked Enterprises: in Asia2010 but not in Emens */
run;
proc freq data=emens_asia;
tables v/missing;
run;

/* Alignment of Emens enterprises with Asia enterprises with employment>0*/
proc sort data=sasrava.Asia10_attive_definitivo3 out=Asia2010_emp;
by codice_f; /*codice_f is Enterprise Tax Code in Asia Business Register*/
where dip10>0;
run;
data emens_asia_emp;
merge Emens_nodup ( rename=( co_dirt_dichiara=codice_f) in=a ) Asia2010_Emp (in=b);
by codice_f;
if a and b then v=1; /* linked Enterprises*/
if a and not b then v=2; /* not linked Enterprises: in Emens but not in Asia2010_Emp */
if b and not a then v=3; /* not linked Enterprises: in Asia2010_Emp but not in Emens */
run;
proc freq data=emens_asia_emp;
tables v/missing;
run;
```


Results

Number of enterprises in Business Register - all enterprises with exactly the same unit of analysis and the same concept definition as those in Emens = 1.376.387

Number enterprises in Asia Business Register = 4.525.155

I_{4.2.2.1} (Enterprise unit, Asia - all enterprises) = (1.376.387 / 4.525.155) * 100 = 30,42%

Number of enterprises in Asia Business Register with employment >0 having exactly the same unit of analysis and the same concept definition as those in Emens) = 1.370.863

Number of enterprises in Business Register - enterprises with employees = 1.563.129

I_{4.2.2.1} (Enterprise unit, Business Register-enterprises with employees) = (1.370.863 / 1.563.129) * 100 = 87,70%

Comments for implementation

- To implement this method it is essential to consider a reference statistical population.
- For the computation, a step of record linkage (join with common key) is requested
- This method is connected to method 1 for comparability of objects (the numerator is the same) and to that for undercoverage indicator.
- The alignment indicator could be useful to point out time alignment issues

4.2.2.2. Alignment of objects - Method 2

Alignment of objects - Method 2
Algorithm $I_{4.2.2.2} = \% \text{ of corresponding aligned objects} = (\text{Number of objects in the statistical population with the same unit of analysis and same concept definition in the source, after harmonization}) / (\text{Number of relevant objects in statistical population}) \times 100$ Objects refer to Municipalities
Sas programme /* Alignment of Emens Municipality (after harmonization with Istat units) */ See /* Comparability of Emens Municipality (after harmonization with Istat units) */ in 4.2.1.2 Sas programme
Results Number of Municipalities in the statistical population with the same unit of analysis and same concept definition in the Emens source, after harmonization = 8.063 Number of Municipalities in the Istat Municipalities Register = 8.094 $I_{4.2.2.2} (\text{Municipality unit}) = (8.063 / 8.094) * 100 = 99,62\%$
Comments for implementation <ul style="list-style-type: none">▪ To implement this method it is essential to consider a reference statistical population.▪ For the computation, a procedure of harmonization and a step of record linkage (join with common key) are requested▪ This method is connected to method 2 for comparability of objects (the numerator is the same)▪ The alignment indicator could be useful to point out time misalignment issues

4.2.3. Integrability of variables – Linking variables

This indicator takes into account the usability of the units identification codes in the administrative source necessary to conduct the integration of information with those of the statistical register or those found in other possible sources. In this sense the linking variable is meaningful for the administrative source quality.

The three methods described in the 4.2 Deliverable are tested.

Method 1: the number of objects with missing value for the linking variable

Method 2: the number of objects for which the value of the linking variable is different from the one used by NSI

Method 3: the number of objects with convertible linking variable to one used by the NSI.

As reported in Table 2.1 and 4.1, in the Emens source there are four possible administrative units: Worker tax position, Employee, Enterprise, Workplace. Each of these units has an identification code which could be used as linking variable. We tested measurement methods on the last three units.

Administrative unit	Linking variable	Id code label in the Emens dataset
Employee	Employee Tax Code	CO_DIRT_CFLAVDIP
Enterprise	Enterprise Tax Code	CO_DIRT_DICHIARA
Workplace	Belfiore Municipality Code	CO_DIRT_CODCOMUNE

The Tax Code is the identification code used to pay tax so it generally has a good quality in administrative data and it is often used as a linking variable among different sources. For people the tax code is a fixed length (16) alphanumeric string. The Enterprises Tax code has a fixed length (16) alphanumeric string, in case of sole proprietors, and a fixed length (11) numeric string, otherwise. It is the common identifier used in Istat to integrate data containing information about companies.

For the Workplace, Belfiore is a code system convertible in the Istat Municipalities codes. It has a fixed length string (4) composed by a letter and three numbers.

The three methods are tested on the linking variables as follows:

Method 1	Employee Tax Code Enterprise Tax Code Belfiore Municipality Code
Method 2	Employee Tax Code Enterprise Tax Code
Method 3	Belfiore Municipality Code

4.2.3.1. Linking variables - Method 1

Linking variables - Method 1
Algorithm $I_{4.2.3.1} = \frac{\text{(Number of objects in the source with missing value for linking variable)}}{\text{(Number of objects in the source)}} \times 100$ <p>Objects refer to records</p>
Sas programme <pre>/*Number of records with missing value for linking variables*/ proc format; value \$miss " "="missing" other="nomissing"; run; proc freq data= sasrava.emens_2010_maggio; table co_dirt_cflavdip co_dirt_dichiara co_dirt_codcomune /missing; format co_dirt_cflavdip \$miss. co_dirt_dichiara \$miss. co_dirt_codcomune \$miss.; run;</pre>
Results Number of records in the SSD source = 13.375.903 Number of records in the SSD source with missing value on Employee Tax Code = 20 $I_{4.2.3.1}$ (Employee Tax Code) = $(20 / 13.375.903) * 100 = 0,00015\%$ Number of records in the SSD source with missing value on Enterprise Tax Code = 0 $I_{4.2.3.1}$ (Enterprise Tax Code) = $(0 / 13.375.903) * 100 = 0\%$ Number of records in the SSD source with missing value on Belfiore Municipality Code = 7.595 $I_{4.2.3.1}$ (Belfiore Municipality Code) = $(7.595 / 13.375.903) * 100 = 0,06\%$
Comments for implementation <ul style="list-style-type: none">▪ In case the linking variable for objects is not a primary key of the source (the same object can be present in more records, as it is our case) the percentage of missing values has to be computed with respect to the number of records, i.e. % of records (not number of objects) with missing value for linking variable.▪ From a computational point of view this method that checks for missing values on the linking variables, uses the same algorithm of the completeness indicator that measures the missing values of each variable of the administrative source.

4.2.3.2. Linking variables - Method 2

The objective is to verify the correspondence of the identifier in the Emens source with that used in Istat. As mentioned above, the Tax code is currently used in Istat for data integration procedure (record linkage). Method 2 aims to detect any errors by checking their syntactical correctness, we tested format and length of the code i.e. alphanumeric format and length of the field equal to 16.

All Employees Tax Codes have complied with the rule on all records. Even for the Employers Tax codes any error has been detected. In this second case, the data editing procedure took into account the fact that the format and the length of the code vary according to the type of enterprise.

Linking variables - Method 2

Algorithm

$I_{4.2.3.2} = \% \text{ of objects with syntactical correctness of the linking variable} = (\text{Number of objects in source with syntactical correct value on linking variable}) / (\text{Number of objects with linking variable in the source}) \times 100$

Object refers to record

Sas programme

/*Creation of variables which indicate the length of employee and enterprise identification codes*/

```
data link_var_m2;
```

```
set sasrava.emens_2010_maggio;
```

```
keep co_dirt_cflavdip co_dirt_dichiara;
```

```
lung_cf_employee = length(co_dirt_cflavdip);
```

```
lung_cf_enterprise=length(co_dirt_dichiara); run;
```

/*Number of records with different values for Employee Tax Code */

```
proc freq data= link_var_m2;
```

```
table lung_cf_employee /missing;
```

```
where co_dirt_cflavdip ne ''; run;
```

/*Number of records with different values for Enterprise Tax Code */

```
proc freq data= link_var_m2;
```

```
table lung_cf_enterprise / missing nopercnt nocum;
```

```
where co_dirt_dichiara ne ''; run;
```

/*Number of records with Employee Tax Code */

See 4.2.3.1 Sas programme

/*Number of records with Enterprise Tax Code */

See 4.2.3.1 Sas programme

Results

Number of records in the Emens source with syntactical incorrect value on Employee Tax Code = 0

Number of records with Employee Tax Code in the Emens source = 13.375.883

$I_{4.2.3.2} (\text{Employee Tax Code}) = (0 / 13.375.883) * 100 = 0\%$

Number of records in the Emens source with syntactical incorrect value on Enterprise Tax Code = 0

Number of records with Enterprise Tax Code in the Emens source = 13.375.903

$I_{4.2.3.2} (\text{Enterprise Tax Code}) = (0 / 13.375.903) * 100 = 0\%$

Comments for implementation

- It is necessary to specify the rules used to check the correctness of the linking variable
- This method coincides with that in the context of Accuracy dimension, measuring the authenticity of the units and in particular checking the syntactical correctness of the identification codes.

4.2.3.3. Linking variables - Method 3

This method is useful when the linking variable in the administrative source is different from that used in the NSI and a conversion table between the two systems is available (foreign key), as it is the case of the Working place with the Municipality identification key.

The calculation of the indicator was carried in two ways: a) by counting records having convertible codes (also all the repetitions of the same Belfiore code); b) by counting distinct convertible Belfiore codes (regardless of the number of times each code is repeated in the source).

Linking variables - Method 3

Algorithm

$I_{4.2.3.3} = \text{\% of objects with correctly convertible linking variable} = (\text{Number of objects in the source with linking variable convertible to one used by the NSI}) / (\text{Number of objects with linking variable in the source}) \times 100$

- a) Objects refer to records
- b) Objects refer to municipalities

Sas programme

```
/*Conversion programme (not reported)*/  
/* Number of records with convertible Municipality code*/  
/*Creation of the variable which indicates records with convertible Municipality code */  
data link_var_m3;  
set sasrava.emens_2010_maggio;  
if cod_mun not in ('999999') then conv=1;  
else conv=0;  
run;  
/*Counting of records with convertible Municipality code */  
proc freq data= link_var_m3;  
table conv/missing;  
run;  
/* Number of Municipality units with convertible Municipality code */  
proc sort data= link_var_m3 out=mun_conv nodupkey;  
by co_dirt_codcomune;  
where conv=1;  
run;
```

Results

Number of records in the Emens source with Municipality code convertible to one used by the NSI = 13.353.608

Number of records with Municipality code in the Emens source = 13.368.308

a) $I_{4.2.3.3}$ (Municipality code - Record) = $(13.353.608 / 13.368.308) * 100 = 99,89\%$

Number of Municipalities in the Emens source with Municipality code convertible to one used by the NSI = 8.506

Number of Municipalities with Municipality code in the Emens source = 8.725

b) $I_{4.2.3.3}$ (Municipality code - Municipality unit) = $(8.506 / 8.725) * 100 = 97,49\%$

Comments for implementation

- For the computation, a step of record linkage (join with foreign key) is requested
- Method 3 measures the percentage of administrative objects with convertible linking variable to one used by the NSI. In our case for implementing it, a conversion table should be read.
- When the same object can be found in more records, as it is our case, it could be useful compute the indicator both with respect to the number of records (with duplicates) and the number of units (without duplicates)

4.2.4. Integrability of variables – Comparability of variables

For testing this method, it is necessary to have a reference source having corresponding variables referred to the same units. As there are no common variables between the Asia BR and the SSD source, it was not possible to proceed to the calculation of this indicator.

4.3. ACCURACY

4.3.1. Accuracy of objects: Authenticity

The Authenticity indicator focuses on the legitimacy of objects in the source. This includes syntactic correctness of the identification key used and the correspondence of the object in the source with the intended object in the real world.

For this indicator WP4 proposed three measurement methods:

Method 1: % of objects with a non-syntactically correct identification key

Method 2: % of objects for which the data source contains information contradictive to information in a reference list for those objects

Method 3: Contact the data source holder to know their % of non-authentic objects in the source

We tested Method 1 using the same measurement method used for Linking variables - Method 2 (see § 4.2.3.2) .

For Method 2, we have no reference sources usable to verify the authenticity of the objects through the comparison of common variables as demographic variables or other stratification variables.

4.3.2. Accuracy of objects: Inconsistent objects

The test of this method requires the definition of logical relations and the possibility to search for objects not satisfying these relations.

As already pointed out, in the Emens source there are more types of objects (Table 2.1) and different types of possible relationships among them (Section 2). Given this structure of the data, the indicator can not be tested because it is not possible to identify inadmissible hierarchical relations among units. A soft rule is defined for dubious objects (see next Section).

4.3.3. Accuracy of objects: Dubious objects

This indicator, as the previous one, is useful for sources with multiple objects. It investigates the relationship between different types of units with the aim of detecting dubious objects identified by the presence of implausible but not necessarily incorrect relations. The indicator provides the percentage of units that must be subjected to more accurate checks and inspections and possibly not considered in the statistical process if it is not possible to interpret the meaning of the relationship.

For the case study, we tested Method 1: % of objects involved in implausible but not necessarily incorrect relations with other objects. In particular we considered the following soft rule:

During the month an employee can not have more than k “attachments” with different employers³.

The value of the parameter k could be, for example, $k^* = 5$.

³ The “attachment” with an employer could be considered as a proxy of “job”.

4.3.3.1. Dubious objects – Method 1

Dubious objects – Method 1

Algorithm

$I_{4.3.3.1} = (\text{Number of objects in the administrative source involved in implausible but not necessarily incorrect relations with other objects}) / (\text{Number of objects in the administrative source}) * 100$

Objects represent Employees

Sas programme

```
/* Programme to count the number of enterprises associated to each employee */
proc summary data= sasrava.Emens_2010_maggio nway;
class co_dirt_cflavdip co_dirt_dichiara;
var co_dirt_anno;
output out = record_cf_cfi_nodup N=; run;
proc summary data= record_cf_cfi_nodup nway;
var co_dirt_anno;
class co_dirt_cflavdip;
output out =N_impresse_cf N=; run;
proc freq data=N_impresse_cf; tables _freq_; run;
```

Results

Frequency distribution of Employees by number of Enterprises associated with them on May 2010

Number of enterprises on May 2010	Employees absolute frequency
1	12.712.004
2	269.496
3	14.708
4	2.506
>= 5	1.283
Total	12.999.997

Number of Employees in the SSD source involved in implausible but not necessarily incorrect relation with Enterprises (“attachments” with more than 5 Enterprises on May 2010) = 1.283

Number of Employees in the SSD source = 12.999.997

$I_{4.3.3.1} = (1.283 / 12.999.997) * 100 = 0,0099\%$

Comments for implementation

- It is necessary to have a good knowledge of data structure and relationships between the units to define suitable soft rules.

4.3.4. Accuracy of variables: Measurement error

The Measurement error indicator looks at the correctness of the value for a variable in the source. As the data collection is carried out by the data source holder, it is not possible to detect possible measurement errors (such as reporting error, registration error, processing error) unless they are marked by the data holder himself.

This indicator has not be calculated for the SSD source.

4.3.5. Accuracy of variables: Inconsistent values

The Inconsistent values indicator looks at the consistency of values for a variable or the consistency of values for combinations of variables in the source. It focuses on the extent to which the values for variables in the source are not internally consistent.

4.3.5.1. Inconsistent values – Method 1

To test Inconsistent values indicator we considered the following variables of the Emens source.

Variable	Type	Name in Emens dataset
Hiring date (day in month)	Numerical	co_dirt_ggassunz
Termination date (day in month) of job contract	Numerical	co_dirt_ggcessaz
Employee age (derived by Tax Code)	Numerical	age_emp
Remuneration	Numerical	co_dirt_imponibile
Number of paid days per month	Numerical	co_dirt_ggretrib
Contractual working time (Full/Part-time)	Categorical	co_dirt_qualif2
Part-time percentage	Numerical	co_dirt_percpartime

A set of checking rules has been applied to SSD source: numerical rules involving one numerical variable and Mixed rules involving two variables

Numerical edit rules		
Num1	Hiring day > 31	(if co_dirt_ggassunz > 31)
Num2	Termination day of job contract > 31	(if co_dirt_ggcessaz > 31)
Num3	Employee age < 15	(if age_emp < 15)
Mixed edit rules		
Mix1	Termination day < Hiring day	(if co_dirt_ggcessaz < co_dirt_ggassunz and co_dirt_ggassunz ne 0 and co_dirt_ggcessaz ne 0)
Mix2	Paid days > 0 and Remuneration = 0	(if co_dirt_ggretrib > 0 and co_dirt_imponibile= 0)
Mix3	Full-time employment and nonzero part-time percentage	(if co_dirt_qualif2 = 'F' and co_dirt_percpartime ne 0)

Inconsistent values – Method 1

Algorithm

$I_{4.3.5.1} = (\text{Number of objects in the administrative source of which values (or combination of values) for variables are involved in non-logical relations}) / (\text{Number of objects in the administrative source}) * 100$

Objects refer to records

Sas programme

/*Creation of variables which detect records involved in non-logical relations with regard to the considered rules*/

```
data inconsistent_val_m1; set sasrava.emens_2010_maggio;
if co_dirt_ggassunz > 31 then Num1 = 1; else Num1 = 0;
if co_dirt_ggcessaz > 31 then Num2 = 1; else Num2 = 0;
birth_year_emp = substr (co_dirt_cflavdip, 7, 2);
birth_year_emp_ok=1900+ birth_year_emp;
age_emp = 2010 - birth_year_emp_ok;
if age_emp < 15 then Num3 = 1; else Num3 = 0;
if co_dirt_ggcessaz < co_dirt_ggassunz then Mix1= 1; else Mix1=0;
if co_dirt_ggretrib > 0 and co_dirt_imponibile = 0 then Mix2 = 1; else Mix2= 0;
if co_dirt_qualif2 = 'F' and co_dirt_percpartime ne 0 then Mix3 = 1; else Mix3=0; run;
```

/*Counting of records involved in non-logical relations with regard to the considered rules*/

```
proc freq data=inconsistent_val_m1;
tables Num1 Num2 Num3 Mix2 Mix3 / missing; run;
proc freq data=inconsistent_val_m1;
tables Mix1/ missing; where co_dirt_ggassunz ne 0 and co_dirt_ggcessaz ne 0; run;
```

Results

Number of records in the Emens source = 13.375.903

Editrules	Number of records in the Emens source of which combinations of values for variables are involved in non-logical relations	$I_{4.3.5.1}$
Num1	0	$(0 / 13.375.903) * 100 = 0\%$
Num2	0	$(0 / 13.375.903) * 100 = 0\%$
Num3	456	$(456 / 13.375.903) * 100 = 0,0034\%$
Mix1	1.365	$(1.365 / 13.375.903) * 100 = 0,01\%$
Mix2	13.838	$(13.838 / 13.375.903) * 100 = 0,10\%$
Mix3	15.047	$(15.047 / 13.375.903) * 100 = 0,11\%$

Comments for implementation

- It is necessary to lay down rules through logical and/or numerical constraints
- Some variables could be involved in more than one rule

4.3.6. Accuracy of variables: Dubious values

The Dubious values indicator checks for the occurrence of implausible values on a variable or on a combinations of variables for an object.

Some soft edit rules have been tested on SSD source.

4.3.6.1. Dubious values – Method 1

We tested method 1 on the following variables and edit rules.

Variable	Type	Name in Emens dataset
Employee age (derived by Tax Code)	Numerical	age_emp
Remuneration	Numerical	co_dirt_imponibile
Number of paid days per month	Numerical	co_dirt_ggretrib

Numerical edit rule		
Num4	Remuneration < 0	(if co_dirt_imponibile < 0)
Num5	Number of paid days per month > 26	(if co_dirt_ggretrib > 26)
Num6	Employee age > 65	(if age_emp > 65)

Dubious values – Method 1

Algorithm

$I_{4.3.6.1} = (\text{Number of objects in the administrative source with values involved in implausible but not necessarily incorrect relations}) / (\text{Number of objects in the administrative source}) * 100$

Objects refer to record

Sas programme

*/*Creation of variables which detect records with dubious values with regard to the considered rules*/*

```
data dubious_values_m1; set sasrava.emens_2010_maggio;
```

```
if co_dirt_imponibile < 0 then Num4 = 1; else Num4 = 0;
```

```
if co_dirt_ggretrib > 26 then Num5 = 1; else Num5 = 0;
```

```
birth_year_emp = substr(co_dirt_cflavdip, 7, 2);
```

```
birth_year_emp_ok=1900+ birth_year_emp;
```

```
age_emp = 2010 - birth_year_emp_ok;
```

```
if age_emp > 65 then Num6 = 1; else Num6 = 0;
```

```
run;
```

*/*Counting of records with dubious values with regard to the considered rules*/*

```
proc freq data=dubious_values_m1; tables Num4 Num5 Num6;
```

```
run;
```

Results

Number of records in the Emens source= 13.375.903

Editrules	Number of records with values involved in implausible but not necessarily incorrect relations	I_{4.3.6.1}
Num4	0	$(0 / 13.375.903) * 100 = 0\%$
Num5	54.114	$(54.114 / 13.375.903) * 100 = 0,40\%$
Num6	49.312	$(49.312 / 13.375.903) * 100 = 0,37\%$

Comments for implementation

- It is necessary to lay down rules through logical and numerical constraints
- Some variables could be involved in more than one rule

4.4. COMPLETENESS

The completeness evaluates the degree to which a data source includes data describing the corresponding set of real-world objects and variables. Completeness indicators for objects focus on coverage issues (undercoverage, overcoverage, selectivity and redundancy) while, for variables, they verify the presence of missing and imputed values.

4.4.1. Completeness of objects: Undercoverage

The Undercoverage indicator considers objects in the reference population missing in the administrative source.

The reference population can be the target population of the source (provider master list) or target statistical population (business register, for example).

In order to test these indicators on the Emens source it is necessary to select the administrative unit for which a reference list is available. The SSD target population (master list) is never available. To compute undercoverage indicator we considered the Employers population compared to Asia BR units as in Section 4.2. about the Integrability dimension quality. As in Section 4.2.2.1. (Alignment of objects - Method 1) we used the subpopulation of the Business Register including only enterprises with employment >0 (the closest population to the administrative one).

In order to find the enterprises of the Business Register missing in the Emens source, a record linkage procedure was necessary using the common key of the Tax code.

4.4.1.1. Undercoverage – Method 1

Undercoverage – Method 1

Algorithm

$I_{4.4.1.1} = (\text{Number of objects of the reference list missing in the administrative source}) / (\text{Number of objects in the reference list}) * 100$

Objects refer to enterprises

Sas programme

```
/* Undercoverage of Emens respect to Asia enterprises with employment>0*/
proc sort data=sasrava.emens_2010_maggio out=Emens_nodup nodupkey;
by co_dirt_dichiara;
run;
proc sort data=sasrava.Asia10_attive_definitivo3 out=Asia2010_emp;
by codice_f; /*codice_f is Enterprise Tax Code in Asia Business Register*/
where dip10>0;
run;
data emens_asia_emp;
merge Emens_nodup ( rename=( co_dirt_dichiara=codice_f) in=a ) Asia2010_Emp (in=b);
by codice_f;
if a and b then v=1; /* linked Enterprises*/
if a and not b then v=2; /* not linked Enterprises: in Emens but not in Asia2010_Emp */
if b and not a then v=3; /* not linked Enterprises: in Asia2010_Emp but not in Emens */
run;
proc freq data=emens_asia_emp;
tables v/missing;
run;
```

Results

Number of enterprises of the Business Register (Asia enterprises with employment>0) missing in the Emens source = 192.226

Number of enterprises of the Business Register (Asia enterprises with employment>0) = 1.563.129

$I_{4.4.1.1} = (192.226 / 1.563.129) * 100 = 12,30\%$

Comments for implementation

- To implement this method it is essential to consider a reference population (master list or target statistical population).
- For the computation, a step of record linkage (join with common key) is requested
- The undercoverage indicator could be useful to point out time issues
- Results obtained could be affected by possible errors in the record linkage procedure in case of low quality of the linking variables (it could be useful to highlight the connection with Integrability dimension indicators)
- From the computational point of view, undercoverage indicators use the same algorithm used for Alignment indicators in Integrability dimension.

4.4.2. Completeness of objects: Overcoverage

This indicator quantifies the percentage of objects in the administrative source not in the reference population (master list or target statistical list).

In evaluating the Overcoverage with respect to a target statistical population, two measurements methods have been defined with the aim of distinguishing the reason of the mismatch:

1. objects in the source not included in the target statistical population due to time misalignment (business demography);
2. objects not belonging to the target statistical population, because excluded from the observation field of the target statistical population.

In testing the overcoverage indicator on the Emens source, it was not possible to compute these two methods, so the computation was made considering the number of objects in the source not found - both not included or not belonging - in the target statistical population.

4.4.2.1. Overcoverage

As for the Undercoverage, also the Overcoverage indicator has been calculated on Enterprise unit with respect to the Asia Business Register subpopulation of enterprises with employment >0 .

The indicator measures the share of enterprises not in the Business Register.

Only in order to give a further element of knowledge of the source, we add that to determine if the small proportion of enterprises (9,11%) not in the Business Register are not included as they do not belong to the BR field of observation, in the SSD source a proxy variable is available: the National collective bargaining agreement code. In general we found that about 30% of employees in SSD have national collective bargaining agreements possibly related to Public Institutions and about 15% related to Forestry Enterprises (out of the BR target population).

Overcoverage
Algorithm $I_{4.4.2.1} = (\text{Number of objects in the administrative source not in the statistical population}) / (\text{Number of objects in the administrative source}) * 100$ <p>Objects refer to enterprises</p>
Sas programme <pre>/*Overcoverage of Emens respect to Asia enterprises with employment>0*/ See 4.4.1.1 Sas programme</pre>
Results <p>Number of Enterprises in the Emens source not in the Business Register (Asia enterprises with employment>0) = 137.458 Number of Enterprises in the Emens source = 1.508.321 $I_{4.4.2.1} = (137.458 / 1.508.321) * 100 = 9,11\%$</p>
Comments for implementation <ul style="list-style-type: none">▪ To implement this method it is essential to consider a reference population (master list or target statistical population).▪ For the computation, a step of record linkage (join with common key) is requested▪ Results obtained could be affected by possible errors in the record linkage procedure in case of low quality of the linking variables (it could be useful to highlight the connection with Integrability dimension indicators)▪ From the computational point of view, overcoverage indicators use the same algorithm used for Comparability indicators in Integrability dimension.

4.4.3. Completeness of objects: Selectivity

The Selectivity indicator looks at the statistical coverage and representativeness of objects in the source with respect to common variables with the target statistical population.

The Selectivity indicator particularly focuses on objects missing in specific strata in a systematic way. This happens, for example, when a data source contains information on a particular subset of the population.

Graphical and analytical measurement methods have been defined for this indicator but it is not possible to test them because Asia Business Register and the Emens source do not contain common stratification variables to compare corresponding frequencies.

The Municipality territorial unit could be considered as a stratification variable. But in the administrative source it is the place where the work is mainly carried out, then the reference population should be the enterprises local units. The previous version of the annual SSD could be used, among other helpful sources, in order to estimate the employment of local units at the municipal level but we have not performed this test because the processing is too complicated and more relevant for the source treatment step rather than for a first quality evaluation.

4.4.4. Completeness of objects: Redundancy

The redundancy indicator analyzes the occurrence of multiple registrations of identical objects in the administrative source. Specifically, the three measurement methods for this indicator explore the presence of duplicate objects in the source with respect to: the identification code (method 1); a selection of variables (method 2); all variables (method 3).

All three methods were calculated on the SSD source.

It should be noted that the mechanism of SSD generation, described in Section 2, produces a redundancy of information compared to the space occupied by the data (computer memory) but the presence of duplicates on the identification codes is not an error. Also with respect to a set of variables, the presence of duplicates is admissible.

In particular, in the case of the Employee tax code, duplications are admissible because of possible changes in the work characteristics during the month or possible multiple jobs. In case of the Employer Tax code, duplications are due to the presence of multiple workers employed in enterprise. Duplications of the identification code of the working place are obviously due to the possible presence of several companies within the same Municipality.

We didn't find any duplicated record for the entire set of variables.

4.4.4.1. Redundancy – Method 1

This method has been tested detecting duplicates for the three Emens identification codes: Employee Tax Code, Enterprise Tax Code and Belfiore Municipality Code.

The procedure used to implement this method counts the number of records without missing value on the identification code and the number of distinct records on the key. The difference between the two quantities provides, in fact, the number of duplicates for the identification code.

Redundancy – Method 1

Algorithm

$I_{4.4.4.1} = (\text{Number of duplicate records for objects in the source} - \text{with the same identification number}) / (\text{Number of objects in the source with identification code}) * 100$

- a) Objects refer to Employees
- b) Objects refer to Enterprises
- c) Objects refer to Municipalities

Sas programme

```
/* a) Employee Tax codes*/  
/* Number of distinct Employee Tax codes */  
proc summary data=sasrava.Emens_2010_maggio nway;  
class co_dirt_cflavdip;  
var co_dirt_anno;  
output out=Code_Emp_distinct N= ; run;  
/* Number of records with Employee Tax code */  
See 4.2.3.1 Sas programme
```

```
/* b) Enterprise Tax codes */  
/* Number of distinct Enterprise Tax codes */  
proc summary data=sasrava.Emens_2010_maggio nway;  
class co_dirt_dichiara;  
var co_dirt_anno;  
output out=Code_Ent_distinct N= ; run;  
/* Number of records with Enterprise Tax code */  
See 4.2.3.1 Sas programme
```

```
/* c) Municipality codes */  
/* Number of distinct Municipality codes */  
proc summary data=sasrava.Emens_2010_maggio nway;  
class co_dirt_codcomune;  
var co_dirt_anno;  
output out=Code_Mun_distinct N= ; run;  
/* Number of records with Municipality code */  
See 4.2.3.1 Sas programme
```

Results

Object	Identification code	Number of records with identification code in Emens	Number of objects in Emens distinct by identification code	Number of duplicated records for objects in Emens	I_{4.4.4.1}
a) Employee	Employee Tax Code	13.375.883	12.999.997	375.886	$(375.886 / 13.375.883) * 100 = 2,81\%$
b) Enterprise	Enterprise Tax Code	13.375.903	1.508.321	11.867.582	$(11.867.582 / 13.375.903) * 100 = 88,72\%$
c) Municipality	Belfiore Municipality Code	13.368.308	8.725	13.359.583	$(13.359.583 / 13.368.308) * 100 = 99,93\%$

4.4.4.2. Redundancy – Method 2

In order to test this method we defined two sets of variables that have some relevance in the context of the source. The variables set 2 is expected to be the identification code of the Worker Tax position unit.

Variables set 1

Variable	Name in Emens dataset
Employee Tax Code	co_dirt_cflavdip
Enterprise Tax Code	co_dirt_dichiara

Variables set 2

Variable	Name in Emens dataset
Employee Tax Code	co_dirt_cflavdip
Enterprise Tax Code	co_dirt_dichiara
Professional status	co_dirt_qualif1
Contractual working time (Full/Part-time)	co_dirt_qualif2
Type of employment contract (Fixed term/Permanent)	co_dirt_qualif3
Type of contribution	co_dirt_tipcontri

Redundancy – Method 2

Algorithm

$I_{4.4.4.2} = (\text{Number of duplicated records for objects in the source - with the same values for a set of variables}) / (\text{Number of records in the source - with no missing value for all variables selected}) * 100$

- a) Objects refer to the attachment Employee-Enterprise (set 1)
- b) Objects refer to the Worker Tax position (set 2)

Sas programme

```
/* Selection and counting of the number of records with no missing value for all set 1 variables */
data set1_no_missing;
set sasrava.Emens_2010_maggio;
if co_dirt_cflavdip ne '' and co_dirt_dichiara ne '' ; run;
/* Number of records distinct by set 1 variables */
proc summary data= set1_no_missing nway;
class co_dirt_cflavdip co_dirt_dichiara;
var co_dirt_anno;
output out=Distinct_set1 N= ; run;
/* Selection and counting of the number of records with no missing value for all set 2 variables */
data set2_no_missing;
set sasrava.Emens_2010_maggio;
if co_dirt_cflavdip ne '' and co_dirt_dichiara ne '' and co_dirt_qualif1 ne '' and co_dirt_qualif2 ne '' and
co_dirt_qualif2 ne '' and co_dirt_tipcontri ne '' ; run;
/* Number of records distinct by set 2 variables */
proc summary data= set2_no_missing nway;
class co_dirt_cflavdip co_dirt_dichiara co_dirt_qualif1 co_dirt_qualif2 co_dirt_qualif3 co_dirt_tipcontri;
var co_dirt_anno;
output out=Distinct_set2 N= ; run;
```

Results

Variables Set	Number of records with no missing values for all selected variables	Number of objects in Emens distinct by variables set	Number of duplicated records in Emens by variables Set	$I_{4.4.4.2}$
Set 1	13.375.883	13.312.534	63.349	$(63.349 / 13.375.883) * 100 = 0,47\%$
Set 2	13.371.702	13.360.551	11.151	$(11.151 / 13.360.551) * 100 = 0,08\%$

Comments for implementation

- The dataset primary key (for definition without duplications) is not always known a priori.

4.4.4.3. Redundancy – Method 3

For this method we consider the duplicate records respect to all variables in the Emens source. The result is that there are no duplicated records.

Redundancy – Method 3
Algorithm $I_{4.4.4.3} = (\text{Number of duplicate objects in Emens with the same values for all variables}) / (\text{Number of objects in Emens}) * 100$ <p>The objects refer to records</p>
Sas programme <pre>/* Programme to calculate the number of duplicates for all variables in the source Emens*/ proc sort data=sasrava.emens_2010_maggio out=emens_2010_maggio_rec_nodupkey; by co_dirt_cflavdip--co_dirt_percpartime; run;</pre>
Results Number of duplicate records in the Emens source with the same values for all variables = 0 Number of records in the Emens source = 13.375.903 $I_{4.4.4.3} = (0 / 13.375.903) * 100 = 0\%$

4.4.5. Completeness of variables: Missing values

This indicator explores the presence of the missing values in the source by using different methods, both graphical and analytical. Method 1 counts the number of objects with missing value for a particular variable, while Method 2 considers objects with all missing values for a selected (limited) number of variables. Method 3 consists in the graphical representation of the frequencies of missing values for each variable in the source.

4.4.5.1. Missing values - Method 1

In order to test this method, frequency analysis for each variable of the source is sufficient. The computation was made by considering the main variables of the SSD source.

Variable	Name in Emens dataset
Employee Tax Code	co_dirt_cflavdip
Enterprise Tax Code	co_dirt_dichiara
Professional status	co_dirt_qualif1
Contractual working time (Full/Part-time)	co_dirt_qualif2
Type of employment contract (Fixed term/Permanent)	co_dirt_qualif3
Type of contribution	co_dirt_tipcontri
Hiring date (day in month)	co_dirt_ggassunz
Job contract termination date (day in month)	co_dirt_ggcessaz
Hiring reason	co_dirt_tpassunz
Job contract termination reason	co_dirt_tpcsessaz
Part-time percentage	co_dirt_percpartime

For categorical variables the missing value is indicated with blank, while for the numerical ones with ‘.’.

For variables: Hiring reason, Job contract termination reason and Part-time percentage, a value is expected only in association with the value of the corresponding filter variable

Hiring reason	Hiring date (day in month) >0	(if co_dirt_ggassunz > 0)
Job contract termination reason	Job contract termination day >0	(if co_dirt_ggcessaz > 0)
Part-time percentage	Contractual working time ≠ F	(if co_dirt_qualif2 ne ‘F’)

Missing values - Method 1

Algorithm

$I_{4.4.5.1} = (\text{Number of objects in the source with a missing value for a particular variable}) / (\text{Number of objects in the source}) * 100$

Objects refer to records

Sas programme

/* Counting of the number of records with a missing value for each variable considered for which the value is always expected*/

```
proc format; value $miss  
" "="missing"  
other="nomissing"; run;
```

```
proc format; value miss_a  
.="missing"  
other="nomissing"; run;
```

```
proc freq data= sasrava.emens_2010_maggio;  
table co_dirt_cflavdip co_dirt_qualif1 co_dirt_qualif2 co_dirt_qualif3 co_dirt_tipcontri co_dirt_dichiara  
co_dirt_ggassunz co_dirt_ggcessaz /missing;  
format  
co_dirt_cflavdip $miss.  
co_dirt_qualif1 $miss.  
co_dirt_qualif2 $miss.  
co_dirt_qualif3 $miss.  
co_dirt_tipcontri $miss.  
co_dirt_dichiara $miss.  
co_dirt_ggassunz miss_a.  
co_dirt_ggcessaz miss_a. ;  
run;
```

/* Counting of the number of records with a missing value for the variable Hiring reason*/

```
proc freq data= sasrava.emens_2010_maggio;  
table co_dirt_tpassunz / missing;  
where co_dirt_ggassunz>0;  
format co_dirt_tpassunz $miss. ;  
run;
```

/* Counting of the number of records with a missing value for the variable Termination reason of the job contract*/

```
proc freq data= sasrava.emens_2010_maggio; table co_dirt_tpcsessaz / missing;  
where co_dirt_ggcessaz>0; format co_dirt_tpcsessaz $miss.; run;
```

/* Counting of the number of records with a missing value for the variable Part-time percentage*/

```
proc freq data= sasrava.emens_2010_maggio;
```

```
table co_dirt_percpartime / missing;
where co_dirt_percpartime co_dirt_qualif2 in ('P' 'M' 'V');
format co_dirt_percpartime miss_a.; run;
```

Results

Variables	Number of records in the source with expected value for the particular variable	Number of records in the source with a missing value for a particular variable	I4.4.5.1
Employee Tax Code	13.375.903	20	$(20 / 13.375.903) * 100 = 0,00015\%$
Enterprise Tax Code	13.375.903	0	$(0 / 13.375.903) * 100 = 0\%$
Professional status	13.375.903	0	$(0 / 13.375.903) * 100 = 0\%$
Contractual working time (Full/Part-time)	13.375.903	4.181	$(4.181 / 13.375.903) * 100 = 0,03\%$
Type of employment contract (Fixed term/Permanent)	13.375.903	4.170	$(4.170 / 13.375.903) * 100 = 0,03\%$
Type of contribution	13.375.903	0	$(0 / 13.375.903) * 100 = 0\%$
Hiring date (day in month)	13.375.903	0	$(0 / 13.375.903) * 100 = 0\%$
Hiring reason	602.028	0	$(0 / 602.028) * 100 = 0\%$
Job contract termination date (day in month)	13.375.903	0	$(0 / 13.375.903) * 100 = 0\%$
Job contract termination reason	466.379	0	$(0 / 466.379) * 100 = 0\%$
Part-time percentage	2.850.226	0	$(0 / 2.850.226) * 100 = 0\%$

Comments for implementation

- For the computation, the first step is to verify for each variable what 'value' in the dataset is used to indicate a missing item and to distinguish items for which a value it is not expected (not applicable).

4.4.5.2. Missing values - Method 2

To test this method we select a set of variables in the Emens source. We used Set 2 already defined for testing Redundancy Method 2 (section 4.4.4.2)

Variables set 2

Variable	Name in Emens dataset
Employee Tax Code	co_dirt_cflavdip
Enterprise Tax Code	co_dirt_dichiara
Professional status	co_dirt_qualif1
Contractual working time (Full/Part-time)	co_dirt_qualif2
Type of employment contract (Fixed term/Permanent)	co_dirt_qualif3
Type of contribution	co_dirt_tipcontri

Missing values - Method 2

Algorithm

$I_{4.4.5.2} = (\text{Number of objects in the source with all values missing for a selected (limited) number of variables}) / (\text{Number of objects in the source}) * 100$

Objects refer to records

Sas programme

```
/*Creation of the variable which detects records in Emens with all values missing for the set of variables */
data missing_m2;
set sasrava.Emens_2010_maggio;
if co_dirt_cflavdip=' ' and co_dirt_dichiara=' ' and co_dirt_qualif1=' ' and co_dirt_qualif2=' ' and co_dirt_qualif3=' ' and
co_dirt_tipcontri=' ' then set2_all_missing=1;
else set2_all_missing=0;
run;
/* Counting of the number of records in Emens with all values missing for the set of variables */
proc freq data= missing_m2;
tables set2_all_missing;
run;
```

Results

Number of records in Emens with all values missing for Set 2 variables = 0

Number of records in Emens = 13.375.903

$I_{4.4.5.2} = (0 / 13.375.903) * 100 = 0\%$

4.4.5.3. Missing values - Method 3

This method has been tested on the Emens source considering two categorical variables: Contractual working time (co_dirt_qualif2) and Type of employment contract (co_dirt_qualif3).

Missing values - Method 3																							
Algorithm	Use of graphical methods to inspect for missing values for variables																						
Sas programme	<pre> /* Creation of vertical bar plots to inspect for missing values for the variables Contractual working time and Type of employment contract*/ proc format; value \$miss_mtre " "="missing"; run; proc gchart data=sasrava.emens_2010_maggio; vbar co_dirt_qualif2 co_dirt_qualif3/ discrete pct; format co_dirt_qualif2_m \$miss_mtre. co_dirt_qualif3 \$miss_mtre.; run; </pre>																						
Results	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Missing values for Contractual Working time</p> <table border="1"> <caption>Missing values for Contractual Working time</caption> <thead> <tr> <th>Category</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>F</td> <td>76.96%</td> </tr> <tr> <td>M</td> <td>1.08%</td> </tr> <tr> <td>P</td> <td>19.23%</td> </tr> <tr> <td>V</td> <td>0.99%</td> </tr> <tr> <td>missing</td> <td>0.03%</td> </tr> </tbody> </table> </div> <div style="text-align: center;"> <p>Missing values for Type of employment contract</p> <table border="1"> <caption>Missing values for Type of employment contract</caption> <thead> <tr> <th>Category</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>D</td> <td>15.29%</td> </tr> <tr> <td>I</td> <td>84.03%</td> </tr> <tr> <td>S</td> <td>0.64%</td> </tr> <tr> <td>missing</td> <td>0.03%</td> </tr> </tbody> </table> </div> </div>	Category	Percentage	F	76.96%	M	1.08%	P	19.23%	V	0.99%	missing	0.03%	Category	Percentage	D	15.29%	I	84.03%	S	0.64%	missing	0.03%
Category	Percentage																						
F	76.96%																						
M	1.08%																						
P	19.23%																						
V	0.99%																						
missing	0.03%																						
Category	Percentage																						
D	15.29%																						
I	84.03%																						
S	0.64%																						
missing	0.03%																						

4.4.6. Completeness of variables: Imputed values

The Imputed values indicator checks for the occurrence of values in the dataset resulting from imputation actions performed by the data source keeper. The information is not available for the calculation of the indicator.

4.5. TIME RELATED

In order to test the robustness of Time related dimension quality indicators, we focused our attention on the Dynamics of objects and Stability of variables indicators.

About the other indicators (Timeliness, Punctuality, Overall time lag and Delay) we did not find any robustness relevant issue.

4.5.1. Dynamics of objects

How much a source is useful to identify changes in the population is an important aspect.

This indicator aims to explore population dynamics of objects by measuring changes in the number of objects between two consecutive moments in time. In order to evaluate the changes (and non-changes) of objects from time t to time $t-1$, the indicator can calculate the new objects (Births t), present at t but not at $t-1$, and the old inactive objects (Deaths t), present at $t-1$ but not at t .

Three methods have been proposed by WP4:

Method 1: The percentage of new objects in the source at t compared to the total number of objects at t

Method 2: The percentage of no longer present objects in the source at t , with respect to $t-1$, compared to the total number of objects at t

Method 3: The percentage of no longer present objects in the source at t , with respect to $t-1$, compared to the total number of objects at $t-1$

For the Emens source we tested the proposed methods considering objects in the months of April 2010 ($t-1$) and May 2010 (t). In particular, we calculated the indicators for a) Enterprises and for b) Employees with a longitudinal perspective.

4.5.1.1. Dynamics of objects – Method 1

Dynamics of objects – Method 1

Algorithm

$I_{4.5.1.1} = (\text{Number of objects at } t \text{ but not at } t-1) / (\text{Number of objects at } t) * 100$

- a) Objects refer to Employees
- b) Objects refer to Enterprises

Sas programme

/* Selection of employee and employer identification codes in the two dataset related to May and April months */

```
data dynamics_obj_may;
set sasrava.emens_2010_maggio;
keep co_dirt_cflavdip co_dirt_dichiara;
run;
data dynamics_obj_april;
set sasrava.emens_2010_aprile;
keep co_dirt_cflavdip co_dirt_dichiara;
run;
```

/*a) Employee unit*/

/* Elimination of duplicated records for Employees*/

```
proc sort data=dynamics_obj_may out=emp_may_nodup nodupkey;
by co_dirt_cflavdip;
where co_dirt_cflavdip ne '';
run;
proc sort data=dynamics_obj_april out=emp_april_nodup nodupkey;
by co_dirt_cflavdip;
where co_dirt_cflavdip ne '';
run;
```

/*Record linkage to detect new, old and persistent Employees from t-1 (April) to t (May)*/

```
data emp_may_april;
merge
emp_may_nodup (in=a)
emp_april_nodup (in=b);
by co_dirt_cflavdip;
if a and b then v=1; /*persistent employees*/
if a and not b then v=2; /*new employees*/
if b and not a then v=3; /*old employees*/
run;
/*Counting of the number of new, old and persistent Employees from t-1 (April) to t (May)*/
proc freq data=emp_may_april; table v; run;
```

/*b) Enterprise unit */

/*Elimination of duplicated records for Enterprises*/

```

proc sort data=dynamics_obj_may out=ent_may_nodup nodupkey;
by co_dirt_dichiara;
where co_dirt_dichiara ne ' ';
run;
proc sort data=dynamics_obj_april out= ent_april_nodup nodupkey;
by co_dirt_dichiara;
where co_dirt_dichiara ne ' ';
run;
/*Record linkage to detect new, old and persistent Enterprises from t-1 (April) to t (May)*/
data ent_may_april;
merge ent_may_nodup (in=a) ent_april_nodup (in=b);
by co_dirt_dichiara;
if a and b then v=1;      /*persistent enterprises*/
if a and not b then v=2; /*new enterprises*/
if b and not a then v=3; /*old enterprises*/
run;
/*Counting of the number of new, old and persistent Enterprises from t-1 (April) to t (May)*/
proc freq data=ent_may_april; table v; run;

```

Results

Number of Employees at t (May) but not at t-1(April) = 392.619

Number of Employees at t (May) = 12.999.997

a) I_{4.5.1.1}(Employee unit) = (392.619 / 12.999.997) * 100 = 3,02%

Number of Enterprises at t (May) but not at t-1(April) = 37.367

Number of Enterprises at t (May) = 1.508.321

b) I_{4.5.1.1}(Enterprise unit) = (37.367 / 1.508.321) * 100 = 2,48%

Comments for implementation

- For the computation, a step of record linkage (join with common key) is requested, so the result can be affected by errors in the linking variable
- Two data files referred to two consecutive periods have to be loaded

4.5.1.2. Dynamics of objects – Method 2

Dynamics of objects – Method 2
Algorithm $I_{4.5.1.2} = (\text{Number of objects at t-1 but not at t}) / (\text{Number of objects at t}) * 100$ a) Objects refer to Employees b) Objects refer to Enterprises
Sas programme See 4.5.1.1 Sas programme
Results Number of Employees at t-1 (April) but not at t (May) = 293.581 Number of Employees at t (May) = 12.999.997 a) $I_{4.5.1.2}$ (Employee unit) = $(293.581/12.999.997)*100 = 2,26\%$ Number of Enterprises at t-1 (April) but not at t (May) = 26.827 Number of Enterprises at t (May) = 1.508.321 b) $I_{4.5.1.2}$ (Enterprise unit) = $(26.827/1.508.321) * 100 = 1,78\%$
Comments for implementation <ul style="list-style-type: none">▪ For the computation, a step of record linkage (join with common key) is requested, so the result can be affected by errors in the linking variable▪ Two data files referred to two consecutive periods have to be loaded

4.5.1.3. Dynamics of objects – Method 3

Dynamics of objects – Method 3
Algorithm $I_{4.5.1.3} = (\text{Number of objects at t-1 but not at t}) / (\text{Number of objects at t-1}) * 100$ <p>a) Objects refer to Employees b) Objects refer to Enterprises</p>
Sas programme See 4.5.1.1 Sas programme
Results Number of Employees at t-1 (April) but not at t (May) = 293.581 Number of Employees at t-1 (April) = 12.900.959 a) $I_{4.5.1.3}$ (Employee unit) = $(293.581 / 12.900.959) * 100 = 2,28\%$ Number of Enterprises at t-1 (April) but not at t (May) = 26.827 Number of Enterprises at t-1 (April) = 1.497.781 b) $I_{4.5.1.3}$ (Enterprise unit) = $(26.827 / 1.497.781) * 100 = 1,79\%$
Comments for implementation <ul style="list-style-type: none">▪ For the computation, a step of record linkage (join with common key) is requested, so the result can be affected by errors in the linking variable▪ Two data files referred to two consecutive periods have to be loaded

4.5.2. Stability of variables

This indicator measures the changes of variables or values over time. Graphical and analytical methods have been proposed to calculate it: Method 1 compares the distributions of the values of a specific variable for persistent objects of the source in two different moments in time by using graphical representations as bar plot and scatter plot; Method 2 measures the percentage of objects with a change value for a particular variable in two different moments in time; Method 3 measures the degree of association between the values of the same variable at two different times by calculating a correlation index (for numerical variables) or the Cramer V association index (for categorical variables).

For the case study, indicator has been tested on two categorical variables: Contractual working time (co_dirt_qualif2) and Type of employment contract (co_dirt_qualif3).

To make the calculation possible, a longitudinal dataset has been created considering for each Employee the values at time t (May) and t-1 (April) corresponding to the main job (in terms of Number of paid days per month).

4.5.2.1. Stability of variables – Method 1

Stability of variables – Method 1

Algorithm

Use statistical data inspection methods to compare the values of specific variables for persistent objects in different deliveries of the source. Graphical methods that can be used are a bar plot and a scatter plot

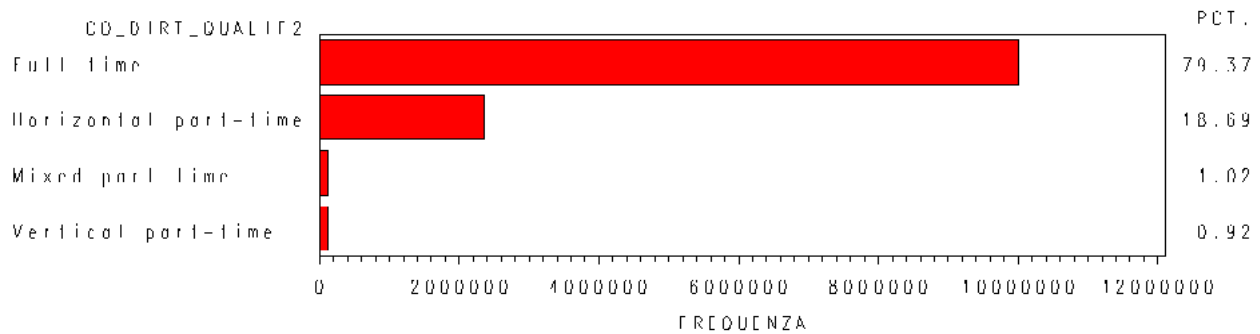
Object refers to employee

Sas programme

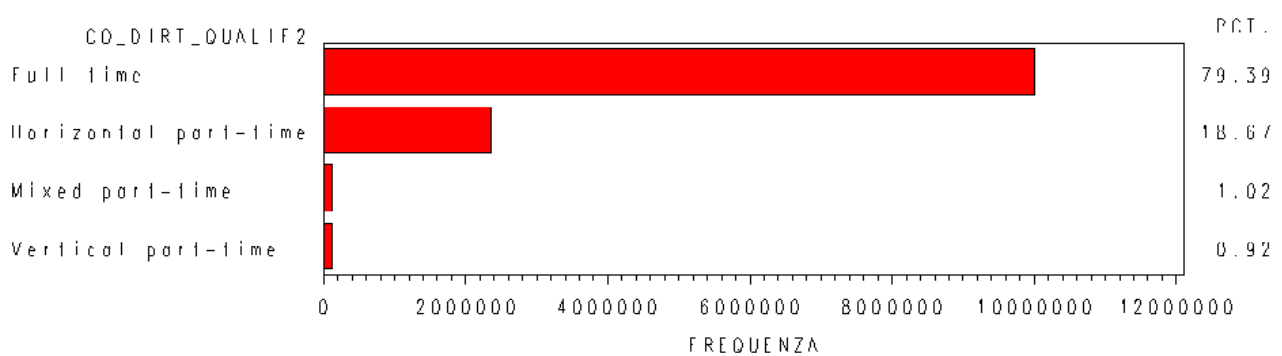
```
/* Selection of the variables considered in the two dataset related to May and April months */
data stability_var_may; set sasrava.Emens_2010_maggio;
keep co_dirt_cflavdip co_dirt_qualif2 co_dirt_qualif3 co_dirt_ggretrib; if co_dirt_cflavdip ne ''; run;
data stability_var_april; set sasrava.Emens_2010_aprile;
keep co_dirt_cflavdip co_dirt_qualif2 co_dirt_qualif3 co_dirt_ggretrib; if co_dirt_cflavdip ne ''; run;
/* Sorting by Number of paid days */
proc sort data= stability_var_may out=emp_may_ord; by co_dirt_cflavdip descending co_dirt_ggretrib; run;
proc sort data= stability_var_april out= emp_april_ord; by co_dirt_cflavdip descending co_dirt_ggretrib; run;
/* Elimination of duplicated records for Employees */
proc sort data = emp_may_ord out = emp_may_nodup nodupkey; by co_dirt_cflavdip; run;
proc sort data = emp_april_ord out = emp_april_nodup nodupkey; by co_dirt_cflavdip; run;
/* Definition of the persistent Employees */
data emp_may_april_pers;
merge
emp_may_nodup (rename=( co_dirt_qualif2= co_dirt_qualif2_m co_dirt_qualif3 = co_dirt_qualif3_m )
drop=co_dirt_ggretrib in=a)
emp_april_nodup (rename=( co_dirt_qualif2= co_dirt_qualif2_a co_dirt_qualif3 = co_dirt_qualif3_a )
drop=co_dirt_ggretrib in=b);
by co_dirt_cflavdip;
if a and b;
run;
/* Creation of bar plots for Working time in t-1 (April) and t (May)*/
proc format; value $ qualifdue 'F' = 'Full-time' 'P' = 'Horizontal part-time' 'V' = 'Vertical part-time' 'M' = 'Mixed
part-time'; run;
proc gchart data=emp_may_april_pers;
hbar co_dirt_qualif2_m co_dirt_qualif2_a / discrete pct;
where co_dirt_qualif2_m ne '' and co_dirt_qualif2_a ne '';
format co_dirt_qualif2_m $qualifdue. co_dirt_qualif2_a $qualifdue.; run;
/* Creation of bar plots for Type of employment contract in t-1 (April) and t (May)*/
proc format; value $ qualiftre 'T' = 'Permanent contract' 'D' = 'Fixed term contract' 'S' = 'Seasonal contract'; run;
proc gchart data=emp_may_april_pers;
hbar co_dirt_qualif3_m co_dirt_qualif3_a / discrete pct;
where co_dirt_qualif3_m ne '' and co_dirt_qualif3_a ne '';
format co_dirt_qualif3_m $qualiftre. co_dirt_qualif3_a $qualiftre.; run;
```

Results

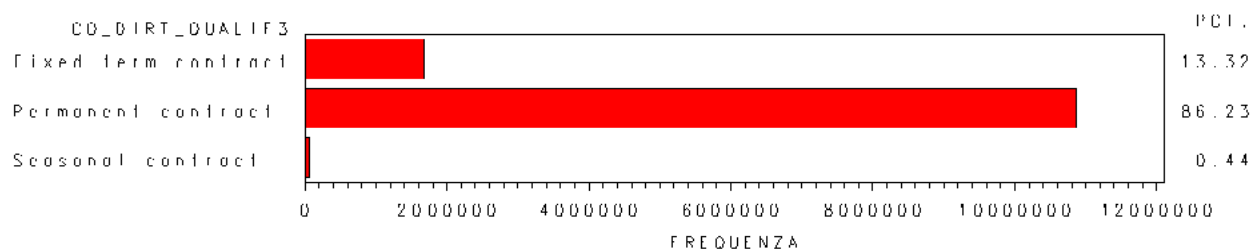
Contractual working time distribution (April 2010)



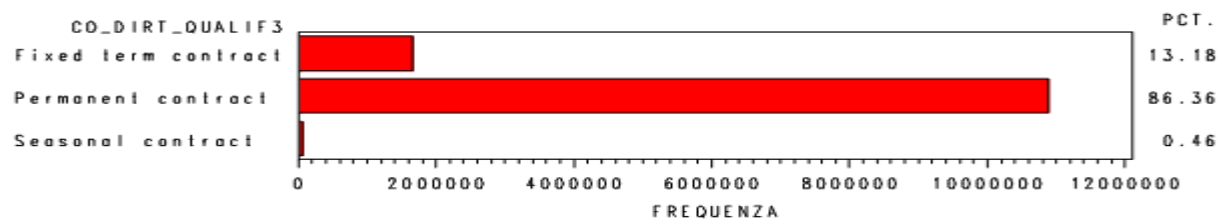
Contractual working time distribution (May 2010)



Type of employment contract distribution (April 2010)



Type of employment contract distribution (May 2010)



Comments for implementation

- For the computation, a step of record linkage (join with common key) is requested, so the result can be affected by errors in the linking variable
- Two data files referred to two consecutive periods have to be loaded

4.5.2.2. Stability of variables – Method 2

Stability of variables – Method 2
Algorithm $I_{4.5.2.2} = \% \text{ of Changes} = (\text{Number of objects with a changed value in } t \text{ for a variable – not missing values}) / (\text{Number of persistent objects with a value filled in } t \text{ for the variable under study}) \times 100\%$ Object refers employee
Sas programme <pre>/* Definition of the persistent Employees */ See /* Definition of the persistent Employees */ in 4.5.2.1 Sas programme /* Creation of the two variables which detect changed values in t (may) for the Working time and Type of employment contract */ data emp_may_april_pers; set emp_may_april_pers; if co_dirt_qualif2_m=co_dirt_qualif2_a then q2_changed = 0; else q2_changed = 1; if co_dirt_qualif3_m=co_dirt_qualif3_a then q3_changed = 0; else q3_changed = 1; run; /* Number of employees with a changed value in t (may) for Working time */ proc freq data= emp_may_april_pers; table q2_changed; where co_dirt_qualif2_m ne '' and co_dirt_qualif2_a ne ''; run; /* Number of employees with a changed value in t (may) for Type of employment contract */ proc freq data = emp_may_april_pers; table q3_changed; where co_dirt_qualif3_m ne '' and co_dirt_qualif3_a ne ''; run;</pre>
Results Number of employees with a changed value at t (may) for Contractual working time – not missing values = 73.553 Number of persistent employees with a value filled in for the variable under study = 12.604.989 $I_{4.5.2.2} \text{ (Contractual working time)} = (73.553/12.604.989)*100 = 0, 58\%$ Number of employees with a changed value at t (may) for Type of employment contract – not missing values = 89.427 Number of persistent employees with a value filled in for the Type of employment contract = 12.604.991 $I_{4.5.2.2} \text{ (Type of employment contract)} = (89.427/12.604.991)*100 = 0,71\%$
Comments for implementation <ul style="list-style-type: none">▪ For the computation, a step of record linkage (join with common key) is requested, so the result can be affected by errors in the linking variable▪ Two data files referred to two consecutive periods have to be loaded

4.5.2.3. Stability of variables – Method 3

Stability of variables – Method 3				
Algorithm				
I _{4.5.2.3} = Cramer V association index for a categorical variable in t-1 and t				
Objects refer to Employees				
Sas programme				
<pre> /* Definition of the persistent Employees */ See /* Definition of the persistent Employees */ in 4.5.2.1 Sas programme /* Calculation of Cramer V association index for Contractual working time in t-1 (April) and t (May)*/ proc freq data= emp_may_april_pers; table co_dirt_qualif2_m*co_dirt_qualif2_a/chisq ; where co_dirt_qualif2_m ne '' and co_dirt_qualif2_a ne ''; run; /* Calculation of Cramer V association index for Type of employment contract in t-1(April) and t (May)*/ proc freq data= emp_may_april_pers; table co_dirt_qualif3_m*co_dirt_qualif3_a/chisq; where co_dirt_qualif3_m ne '' and co_dirt_qualif3_a ne ''; run; </pre>				
Results				
	Statistica	DF	Valore	Prob
I	Chi quadrato	9	35577882	<.0001
	Chi quadrato rapp verosim	9	13858937	<.0001
	Chi quadrato MH	1	12196077	<.0001
	Coefficiente Phi		1.68004	
	Coefficiente di contingenza		0.85930	
	V di Cramer		0.96997	
I _{4.5.2.3} (Contractual working time) = 0,96997				
	Statistica	DF	Valore	Prob
+	Chi quadrato	4	22895397	<.0001
	Chi quadrato rapp verosim	4	9558398	<.0001
	Chi quadrato MH	1	11743065	<.0001
	Coefficiente Phi		1.34773	
	Coefficiente di contingenza		0.80308	
	V di Cramer		0.95299	
I _{4.5.2.3} (Type of employment contract) = 0,95299				
Comments for implementation				
<ul style="list-style-type: none"> ▪ For the computation, a step of record linkage (join with common key) is requested, so the result can be affected by errors in the linking variable. ▪ Two data files referred to two consecutive periods have to be loaded. 				

Conclusions

SSD source, called Emens, produced by the Italian Institute of Social Security (Inps) is a very interesting but complex administrative source for the large amount of data and for the presence of multiple types of units connected to each other (Linked Employer Employee Data).

SSD reacted well to the test, measurement methods have been successfully applied, so quality indicators reached the objectives.

Due to the huge amount of data SAS statistical package has been used.

For the Quality Report Card implementation we have highlighted some general comments.

Quality dimensions are strictly interrelated with each other, it could be useful to highlight these connections.

We suggest to define a hierarchical chain for the indicators computation: Technical checks, Integrability, Accuracy, Completeness, Time-related dimension.

References

Costanzo L., Di Bella G., Hargreaves E., Pereira H., Rodrigues S. (2011), An Overview of the Use of Administrative Data for Business Statistics in Europe, 58th World Statistics Congress of the International Statistical Institute, Dublin, August 21 – 26, 2011. <http://essnet.admindata.eu/>

Daas, P., Ossen, S., Vis-Visschers, R., Arends-Tóth, J. (2009) Checklist for the Quality evaluation of Administrative Data Sources. Discussion paper 09042, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

<http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/Checklist%20for%20the%20quality%20evaluation%20of%20administrative%20d.pdf>

Daas, P., Ossen, S., Tennekes, M., Zhang, L-C., Hendriks, C., Foldal Haugen, K., Bernardi, A., Cerroni, F., Laitila, T., Wallgren, A., Wallgren, B. (2011a) List of quality groups and indicators identified for administrative data sources. First deliverable of WP4 of the BLUE Enterprise and Trade Statistics project, March 10.

<http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.1.pdf>

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. (2011b) Report on methods preferred for the quality indicators of administrative data sources. Second deliverable of WP4 of the BLUE Enterprise and Trade Statistics project, September 28.

<http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf>

Frost, J.M., Green, S., Pereira, H., Rodrigues, S., Chumbau, A., Mendes, J. (2010) Development of quality indicators for business statistics involving administrative data. Paper presented at the Q2010 European Conference on Quality in Official Statistics. Helsinki, Finland.

Appendix C: Evaluation results of Statistics Sweden

EUROPEAN COMMISSION
RESEARCH DIRECTORATE-GENERAL



BLUE-Enterprise and Trade Statistics
BLUE-ETS

SP1-Cooperation-Collaborative Project
Small or medium-scale focused research project

FP7-SSH-2009-A
Grant Agreement Number 244767
SSH-CT-2010-244767

Deliverable 8.2-Sweden

Title: Quality Assessment of Administrative Data
Data Source Quality

Authors: Thomas Laitila, Daniel Lennartsson, Richard Nilsson, Anders Wallgren,
Britt Wallgren (SCB)

Version: First draft (0.1a)

DATE 01-02-2013

CONFIDENTIAL

Quality Assessment of Administrative Data Applications of a system of indicators

Summary:

This report contains a presentation of three applications of the quality indicators proposed in Laitila et al. (2012). The indicators are divided into four sets and suggested to be evaluated in sequence. In each application obtained measurements of the four sets of indicators are presented, followed by a summary in a quality report card. An example of the working process is given in Laitila et al. (2012). Overall the system of indicators was found useful for register quality evaluation. By the construction of the indicators, the evaluation process did not only shed light on the quality of the registers evaluated, but also on the quality of the data sources used in the evaluation process.

Index

1.	Introduction	122
2.	Quality Indicators.....	123
	Metadata – Information from the Administrative Authority	123
	Analysis and Data Editing of the Source	123
	Integrate the Source with the Base Register	124
	Integrate the Source with Surveys with Similar Variables	124
3.	Application I - Income Statements.....	125
3.A	Metadata.....	125
3.B	Analysis and Data Editing of the Source	127
3.C	Integrate the Source with the Base Register	128
3.D	Integrate the Source with Surveys with Similar Variables	130
3.E	Quality Report Card for Income Statements.....	131
4.	Application II – VAT-register.....	133
4.A	Metadata.....	133
4.B	Analysis and Data Editing of the Source	135
4.C	Integrate the Source with the Base Register	136
4.D	Integrate the Source with Surveys with Similar Variables	137
4.E	Quality Report Card for VAT returns	138
5.	Application III – Yearly Tax Returns from Enterprises (SRU).....	140
5.A	Metadata.....	140
5.B	Analysis and Data Editing of the Source	142
5.C	Integrate the Source with the Base Register – the Business Register	143
5.D	Integrate the Source with Surveys with Similar Variables	144
5.E	Quality Report Card for Yearly Tax Returns.....	146
6.	Discussion	148
	References.....	149

1. Introduction

National statistical agencies put increasing efforts in using administrative registers for the production of official statistics. Several motives drive these efforts in building systems for register based surveys, where cost reductions and reduced respondent burdens are two major arguments. Other motives are addressing quality aspects of statistics (e.g. Wallgren and Wallgren, 2007, p. 13).

The use of administrative registers in statistics production is not new and there is a long tradition in using register data for improved estimation in sample surveys. Register information may be used in the sampling design and/or in the estimation stage of a survey (e.g. Särndal et al. 1992; Lehtonen and Veijanen, 1999). Of special importance is the growing literature on adjustment for non-response using register information (e.g. Särndal and Lundström, 2005; Kott and Chang, 2010).

The use of data in administrative registers can be a substitute for using sample survey data, but different data sources should be treated as complements, increasing the options for finding and developing good survey designs. It is therefore appropriate to view quality measurements of an administrative data source in relation to the operations of the national statistical agency. Restricting quality judgments to aspects of using an administrative register as the main data source for the production of official statistics provides a context within which appropriate quality indicators can be defined. However, this only gives a first start for deriving a system of indicators of administrative data quality, as an administrative register can be used for several different purposes with potentially different quality requirements (e.g. Eurostat, 2003; Daas et al, 2008; Laitila et al, 2011).

This view on the quality of administrative data is taken by Laitila et al. (2011) for a development of an indicator system for the measurement of the quality of an administrative register. Laitila et al. (2011) provide a framework for identifying valid quality indicators when judging the quality of a register for the purpose of statistics production. They divide quality into three concepts: Output quality, Production Process Quality, and Data Source Quality (or Input quality as used below). For the user of statistics, the output quality of statistics produced is of concern. In judging the quality of a register it is of interest to obtain indicators useful for interpreting the quality of statistics derived from the register. The production process quality concept concerns potential improvements of the production process and the quality of statistics delivered if the register is used.

Output quality and production process quality concern aspects for the actual usage of a register for statistics production. In addition, it is also meaningful, having a statistical perspective, to evaluate the register with respect to the register holder's purpose of the register, the input quality concept. One reason is the potential usage of a register in new register based surveys.

The system is reported and described in Laitila et al. (2012), which also includes a detailed example of a working process in which the indicators are applied and measured. The present report contains a summary of the application in Laitila et al. (2012) and applications of the indicators to two additional administrative data sources: the Annual Company Reports (SRU) and the Value Added Tax Register (VAT-register).

The indicators used in Laitila et al. (2012) are defined in the next section. Sections 3 – 4 presents applications of these indicators to three registers: the Income Statement register, the VAT-register and the SRU register, respectively. At the end of each section the indicators are summarized in the form of a Quality Report Card. A discussion of the results are presented in the final section.

2. Quality Indicators

How should an administrative register or source be analysed to judge its output quality, input data quality and production process quality? We recommend a work process with four steps with evaluation of different sets of indicators. A background to and explanation of the suggested work process and the indicators are given in Laitila et al. (2012). The four sets of indicators are summarized in the following four charts.

Metadata – Information from the Administrative Authority

Chart A: Indicators of output and input data quality – relevance

Indicator	Quality factor	Description
A1	Relevance of population	Definition of the administrative object set. Which administrative rules determine which objects are included? Is this set suitable as statistical population?
A2	Relevance of units	Definition of the administrative units. Are these units suitable as statistical units?
A3	Relevant keys	Are there primary keys and foreign keys in the source that are suitable for micro integration within the NSI?
A4	Relevance of variables	Definitions of the administrative variables. Are these variables suitable as statistical variables?
A5	Relevance of reference time	Are reference times suitable for statistical usage? What rules for accruing accounting data between months and years are used?
A6	Study domains	Can the units be allocated between relevant study domains? Are there variables describing domains in the source or can the units be linked with domain variables in the Business Register?
A7	Comprehensiveness	Does the source contain a small/large part of an intended population? Does the source contain few/many statistically interesting variables? Can a small/large number of existing surveys benefit from the administrative source?
A8	Updates, delivery time and punctuality	How often and at what time points is the administrative register updated? Time for delivery of the source from register holder to the NSI. Difference in time between delivery and agreed delivery time.
A9	Comparability over time	Extent of changes in the content of the administrative register over time

Analysis and Data Editing of the Source

Chart B. Indicators of output and input data quality – accuracy

Indicator	Quality factor	Description
B1	Primary key	Fraction of units with usable identities. The primary key should have correct format and reasonable values.
B2	Foreign keys	Fraction of units with usable foreign keys. Foreign keys should have correct format and reasonable values.
B3	Missing values	Fraction of missing values for the statistically interesting variables.
B4	Wrong values	Fraction of wrong or unreasonable values for the statistically interesting variables.
B5	Quality of preliminary data	Fraction of records corrected by the taxpayers. Estimates based on preliminary data are compared with estimates based on final data.

Integrate the Source with the Base Register

Chart C: Indicators on output and input data quality – accuracy

Indicator	Quality factor	Description
C1	Undercoverage in BR	Fraction of units: There are enterprises/units that have been active during the reference period but are missing in the BR or are coded as inactive in the BR.
C2	Overcoverage in BR	Fraction of units: Enterprises/units are coded as active in the BR and belong to a category that is covered by the source, but they have no reported activity in the source.
C3	Undercoverage in the source	Fraction of units: There are enterprises/units that have been active during the reference period according to the BR but are missing in the source.
C4	Overcoverage in the source	Fraction of units: There are units in the source that belong to a category, or seem to belong to a category, that is not statistically relevant.
C5	Can the source improve BR?	Here a more thorough analysis is required depending on the character of the source. The quality improvements should be measured.

Integrate the Source with Surveys with Similar Variables

Chart D. Indicators on input data and production process quality

Indicator	Quality factor	Description						
D1	Is the source good or bad?	<table style="width: 100%; border: none;"> <tr> <td style="width: 30%; border: none;">a) Compare populations</td> <td style="width: 40%; border: none;">Measures production process quality</td> </tr> <tr> <td style="border: none;">b) Compare units</td> <td style="border: none;"></td> </tr> <tr> <td style="border: none;">c) Compare variables</td> <td style="border: none;"></td> </tr> </table>	a) Compare populations	Measures production process quality	b) Compare units		c) Compare variables	
a) Compare populations	Measures production process quality							
b) Compare units								
c) Compare variables								
D2	Is the production system good or bad?	<table style="width: 100%; border: none;"> <tr> <td style="width: 30%; border: none;">a) Compare populations</td> <td style="width: 40%; border: none;">Measures production process quality</td> </tr> <tr> <td style="border: none;">b) Compare units</td> <td style="border: none;"></td> </tr> <tr> <td style="border: none;">c) Compare variables</td> <td style="border: none;"></td> </tr> </table>	a) Compare populations	Measures production process quality	b) Compare units		c) Compare variables	
a) Compare populations	Measures production process quality							
b) Compare units								
c) Compare variables								
D3	Can the source improve other surveys?	<table style="width: 100%; border: none;"> <tr> <td style="width: 30%; border: none;">a) Will population be better?</td> <td style="width: 40%; border: none;">Measures production process quality</td> </tr> <tr> <td style="border: none;">b) Will units be better?</td> <td style="border: none;"></td> </tr> <tr> <td style="border: none;">c) Will variables be better?</td> <td style="border: none;"></td> </tr> </table>	a) Will population be better?	Measures production process quality	b) Will units be better?		c) Will variables be better?	
a) Will population be better?	Measures production process quality							
b) Will units be better?								
c) Will variables be better?								
D4	Can the source be combined with other sources?	<table style="width: 100%; border: none;"> <tr> <td style="width: 30%; border: none;">a) Will population be better?</td> <td style="width: 40%; border: none;">Measures input data quality</td> </tr> <tr> <td style="border: none;">b) Will units be better?</td> <td style="border: none;"></td> </tr> <tr> <td style="border: none;">c) Will variables be better?</td> <td style="border: none;"></td> </tr> </table>	a) Will population be better?	Measures input data quality	b) Will units be better?		c) Will variables be better?	
a) Will population be better?	Measures input data quality							
b) Will units be better?								
c) Will variables be better?								

3. Application I - Income Statements

Every month each employer in Sweden deducts preliminary tax from wages and salaries paid. This preliminary tax is transferred to the Tax Board and after each fiscal year each employer sends an Income Statement both to each employee and to the Tax Board. These yearly Income Statements provide information on wages and salaries, preliminary tax and benefits regarding each employee. About 60 % of total taxes in Sweden are covered by this system. The Tax Board receives most of this data via Internet or as data files submitted by the employers. The quality of this administrative data source will be assessed with the indicators listed in section 4.

3.A Metadata

Chart 3A. Indicators of output and input data quality of Income Statements – relevance

Indicator	Quality factor	Description
A1	Relevance of population	<p>Definition of the administrative object set. Which rules determine which objects are included? Is this set suitable as statistical population?</p> <p>There are different kinds of Income Statements for enterprises that are employers and households that are employers. This distinction is important for the National Accounts.</p> <p>Jobs as employed are together with jobs as self-employed the population of jobs during a calendar year. The population of self-employed are in other administrative sources.</p> <p>The source contains information on four administrative object sets – jobs as employed, employed persons, enterprises that are employers and establishments/local units where the employed persons work. All these object sets are suitable and relevant as statistical populations or important parts of statistical populations.</p> <p>A1 = The source contains four kinds of relevant populations or important subpopulations</p>
A2	Relevance of units	<p>Definition of the administrative units. Are these suitable as statistical units?</p> <p>The source contains information on four administrative units – jobs, persons, enterprises and establishments. All these units are suitable and relevant as statistical units.</p> <p>A2 = The source contains four kinds of relevant units</p>
A3	Relevant keys	<p>Are there primary keys and foreign keys in the source that are suitable for micro integration within the NSI?</p> <p>Three important keys are combined in this source:</p> <ul style="list-style-type: none"> – Identity number of the employer, this is the Business Identity Number that is used in the Business Register at Statistics Sweden. – Personal Identity Number of the employee, this is the identity that is used in the Population Register at Statistics Sweden. – Local unit number, this is an identity number that is used by Statistics Sweden to create the register of all establishments or local units. <p>The fact that these three identities are combined in the same source is very important as it will be possible to link records for three kinds of statistical units – person, enterprise and establishment.</p> <p>A3 = The source contains three very relevant keys</p>

A4	Relevance of variables	Definitions of the administrative variables. Are these variables suitable as statistical variables?
<p>Gross salary in box 11 on the tax form plus Benefits corresponds to the definitions used by the National Accounts. Data concerning employer-provided cars can also be used by the National Accounts.</p> <p>A4 = The variables in the source are relevant for statistical purposes</p>		
A5	Relevance of reference time	Are reference times suitable for statistical usage? What rules for accruing accounting data between months and years are used?
<p>The Income Statements give information on wages and salaries paid to the employees during the calendar year. As a rule this is the same as the period when the work was performed but some payments during January can refer to work during the previous year. However, this definition where reference time is defined when wages are paid is in accordance with the needs of the National Accounts. The reference time can in the future be defined as month when monthly Income Statements will be introduced. This will increase the statistical value of this source.</p> <p>A5 = Reference time in the source is relevant for yearly statistics</p>		
A6	Study domains	Can the units be allocated between relevant study domains?
<p>The Income Statements can be linked to the Population Register and the Business Register. All classification variables in these registers or in registers that can be linked with these base registers can be used to define study domains.</p> <p>A6 = Many kinds of study domains are possible for the source</p>		
A7	Comprehensiveness	Does the source contain a large part of an intended population and many statistically interesting variables? Can many surveys benefit from the source?
<p>The source covers all employees and all employers. The source contains a few but very important economic variables and is used for many different surveys, regarding persons, enterprises and establishments.</p> <p>A7 = The source is comprehensive</p>		
A8	Updates, delivery time and punctuality	How often and at what time points is the administrative register updated? Time for delivery of the source from register holder to the NSI. Difference in time between delivery and agreed delivery time.
<p>The source is yearly (may become monthly in the future). Income Statements are delivered to the Tax Board during January, but corrections are made during the whole year. Preliminary statistics can be produced before the summer and final estimates during the autumn.</p> <p>A8 = The source is only yearly and final data are rather late</p>		
A9	Comparability over time	Extent of changes in the content of the administrative register over time
<p>There are no changes regarding populations and variables that give rise to statistical problems.</p> <p>A9 = Comparability over time is good</p>		

The relevance of this source is very high. This data source is necessary for the Employment Register that is a part of the register-based census. The Income Statements are also the best source for statistics on gross wages and are used by the yearly National Accounts. As three identities are combined in the Income Statements this source is a very important part of Statistics Sweden's production system that makes it possible to link records from many different sources with each other.

3.B Analysis and Data Editing of the Source

Chart 3B. Indicators of output and input data quality of Income Statements – accuracy

Indicator	Quality factor	Description
B1	Primary key	<p>Fraction of units with usable identities. The primary key should have correct format and reasonable values.</p> <p>Job Identity Number usable = Both Personal Identity Number (PIN) and Enterprise Identity Number (BIN) usable: 5 031 512 employed persons had 7 132 332 jobs, only 3 107 had PIN that was not usable.</p> <p>Establishment Identity Number usable: 190 701 or 6.4 % of all Income Statements from Income Statements from enterprises with more than one establishment have missing establishment identities. This is regarded as a serious problem and after a register maintenance survey to about 4 400 of these enterprises the establishment identity on 188 962 Income Statements were corrected/changed.</p> <p>B1 = One serious problem was found regarding establishment identities</p>
B2	Foreign keys	<p>Fraction of units with usable foreign keys. Foreign keys should have correct format and reasonable values.</p> <p>Link to the Population Register – PIN usable: Of employed persons 5 028 405 or 99.94 % had usable PIN, 3 107 had not usable PIN. Link to Business Register – BIN usable: All</p> <p>B2 = 99.94 % of employed persons have usable PIN, the key to the Population Register, 100 % of enterprises that are employers have usable BIN, the key to the Business Register.</p>
B3	Missing values	<p>Fraction of missing values for the statistically interesting variables.</p> <p>Employment time defined as the month from and month up to: 0.06 % values are missing.</p> <p>B3 = Small problem</p>
B4	Wrong values	<p>Fraction of wrong or unreasonable values for the statistically interesting variables.</p> <p>Employment time defined as the <i>month from</i> and <i>month up to</i>: Many employers answer from January up to December even if the actual work was done during a shorter period. The aggregate wage can be small but the employment period can be “long”, which would indicate measurement errors.</p> <p>Work site or establishment identity can be erroneous, see B1 above.</p> <p>B4 = Measurement errors in employment time</p>
B5	Quality of preliminary data	<p>Fraction of records corrected by the taxpayers. Estimates based on preliminary data are compared with estimates based on final data.</p> <p>Income Statements are corrected by employers and this causes delay. Preliminary and final estimates were compared, and it was decided that early estimates based on data that are available during September should be used instead of final data that are available in December.</p> <p>B5 = After an extensive study it was decided that preliminary data could be used.</p>

On the whole accuracy is good, but Input data quality is not good enough for the establishment identity numbers. However, after a register maintenance survey, where questionnaires are sent out to more than 4 000 enterprises, the quality of this variable has sufficient quality.

3.C Integrate the Source with the Base Register

Chart 3C. Indicators on output and input data quality of Income Statements – accuracy

Indicator	Quality factor	Description
C1	Undercoverage in the base register	<p>Fraction of units: There are units that have been active during the reference period but are missing in the Population Register and/or missing or coded as inactive in the Business Register.</p> <p>In the Population Register only persons registered as permanently living in Sweden are included. However, foreigners studying or working in Sweden can be registered as temporary staying in Sweden and they get a special kind of identity number. Those who work and pay tax in Sweden are found in the Income Statements. In all 57 905 foreigners were found in the register of all Income Statements that were not found in the Population Register. The fraction of undercoverage among all persons in the Population Register is 0.6 %. The fraction of undercoverage among the population of all employed persons in the Employment Register is 1.4 %.</p> <p>In the Business Register there is a variable that distinguishes between enterprises that are active or not active as employers. In the version of the Business Register that is used for yearly statistics there were 315 380 enterprises that were classified as active employers during one calendar year. According to the Income Statements there were 31 393 enterprises more that were active as employers during the year in question. The fraction of enterprises in the register of active employers in the Business Register that comprise undercoverage is thus 10 % according to the Income Statements.</p> <p>C1 = The undercoverage in the Population Register due to foreigners working in Sweden is 0.6 %. The undercoverage in the Employment Register due to foreigners working in Sweden is 1.4 %. The undercoverage in the Business Register's category of active employers is 10 %.</p>
C2	Overcoverage in base register	<p>Fraction of units: Enterprises are coded as active in the BR and belong to a category that is covered by the source, but has no activity in the source.</p> <p>As all persons are not necessarily employed, this source is not applicable for analysing overcoverage in the Population Register.</p> <p>In the version of the Business Register that is used for yearly statistics there were 315 380 enterprises that were classified as active employers during one calendar year. Out of these 11 301 enterprises were not active according to the Income Statements – they had delivered no Income Statements for the year in question. The fraction of enterprises in the register of active employers in the Business Register that comprise overcoverage according to the Income Statements is 4 %.</p> <p>C2 = The overcoverage in the Business Register's category of active employers is 4 %.</p>
C3	Undercoverage in the source	<p>Fraction of units: There are enterprises/units that have been active during the reference period according to the BR but are missing in the source.</p> <p>The source should contain all units in the intended populations but black work is not included. The Tax Board estimates wages for black work in the Tax Statistical Yearbook of Sweden.</p> <p>C3 = Black work is a problem</p>
C4	Overcoverage in the source	<p>Fraction of units: There are units in the source that belong to a category, or seem to belong to a category, that is not statistically relevant.</p> <p>C4 = No problem</p>
C5	Can the source improve a base register?	<p>Here a more thorough analysis is required depending on the character of the source. The quality improvements should be measured.</p> <p>Comparisons with the Income Statements show that both the coverage of the Population Register</p>

and the Business Register should be improved. However the Income Statements should not be used for these improvements, the Population Register should be improved with data from the Tax Board and the Business Register should be improved with the monthly reports from employers that today are used for the Quarterly Gross Pay survey that are available much earlier.

C5 = Income Statements should not be used to improve base registers.

With the Income Statements it was possible to find important quality flaws in both the Population Register and the Business Register. Both these base registers suffer from undercoverage and the Business Register also suffers from overcoverage. The production process quality of the register with Income Statements is thus very high.

3.D Integrate the Source with Surveys with Similar Variables

Chart 3D. Indicators on input data and production process quality of Income Statements

Indicator	Quality factor	Description	
D1	Is the source good or bad?	a) Compare populations b) Compare units c) Compare variables	Measures production process quality
<p>When compared with the LFS, the QGP and SBS surveys the population, units and variables in the Income Statements were found to be without quality flaws except that black work is not covered in administrative sources as Income Statements.</p> <p>D1 = The register of Income Statements is a good source except that black work is not covered.</p>			
D2	Is the production system good or bad?	a) Compare populations b) Compare units c) Compare variables	Measures production process quality
<p>Many errors were found in the LFS, the QGP survey and the SBS survey after comparisons with Income Statements. Coverage errors in the LFS and SBS were found. Different enterprise units are used in different surveys and in surveys from different periods. The variables Sector and NACE were not consistent between different surveys.</p> <p>D2 = Many errors were found in the LFS, the QGP survey and the SBS survey.</p>			
D3	Can the source improve other surveys?	a) Will population be better? b) Will units be better? c) Will variables be better?	Measures production process quality
<p>Many errors were found thanks to the IS in the LFS and the QGP survey, but IS and its aggregated version YGP are too late to be used to improve these surveys. But the YGP can be used to improve the SBS.</p> <p>D3 = The quality of the SBS survey can be improved by selective editing and imputation models.</p>			
D4	Can the source be combined with other sources?	a) Will population be better? b) Will units be better? c) Will variables be better?	Measures input data quality
<p>Income Statements are used for creating some of the variables in the Income Register. IS must be combined with other sources to give a full picture of disposable income.</p> <p>Income Statements alone does not give a complete picture of the economically active population, but if Income Statements are combined with Yearly Income Declarations for enterprises it is possible to cover both employed and self-employed persons. This combination is the basis for the Employment Register.</p> <p>In the future it will be possible to use monthly Income Statements for the Labour Force Survey.</p> <p>D4 = IS can be combined with other sources. IS is used today for some very important surveys.</p>			

Both the input data quality and the production process quality of the Income Statements and its aggregated version YGP is very high.

3.E Quality Report Card for Income Statements

Chart 3E-A. Information from the Administrative Authority – Relevance

Indicator	Quality factor	Description
A1	Relevance of population	Four kinds of relevant populations or subpopulations
A2	Relevance of units	Four kinds of relevant units
A3	Relevant keys	Three very relevant keys
A4	Relevance of variables	The variables are relevant for statistical purposes
A5	Relevance of reference time	Reference time is relevant for yearly statistics
A6	Study domains	All kinds of study domains are possible
A7	Comprehensiveness	The source is comprehensive
A8	Updates, delivery, punctuality	Only yearly and final data are rather late
A9	Comparability over time	Comparability over time is good

Summary: The relevance of this source is very high.

Chart 3E-B. Analysis and Data Editing of the Source – Accuracy

Indicator	Quality factor	Description
B1	Primary key	One serious problem regarding establishment identities
B2	Foreign keys	Small problem
B3	Missing values	Small problem
B4	Wrong values	Measurement errors in employment time
B5	Quality of preliminary data	It was decided that preliminary data could be used

Summary: Quality is good with the exception of employment time and establishment identities.

Chart 3E-C. Integrate the Source with the Base Register – Accuracy

Indicator	Quality factor	Description
C1	Undercoverage in base register, fraction of units	The undercoverage in the Population Register is 0.6 %, in the Employment Register it is 1.4 % and in the Business Register it is 10 %.
C2	Overcoverage in BR, fraction of units	The overcoverage in the Business Register is 4 %.
C3	Undercoverage in the source	Black work is a problem
C4	Overcoverage in the source	No problem
C5	Can the source improve base register?	IS should not be used to improve base registers

Summary: The production process quality of IS is very high with the exception for black work.

Chart 3E-D. Integrate with Surveys with Similar Variables – Production process quality

Indicator	Quality factor	Description
D1	Is the source good or bad?	The source is good
D2	Is the production system good or bad?	Many important errors were found
D3	Can the source improve other surveys?	The SBS survey can be improved
D4	Can the source be combined with other sources?	The IS source is used in combination with other sources in some very important surveys

Summary: The production process quality and input data quality of IS is very high.

4. Application II – VAT-register

Each month businesses registered for VAT are reporting VAT to the Tax Board. These VAT-statements are delivered to Statistics Sweden. Smaller businesses are reporting VAT to the Tax Board in their income-tax returns and this information is also delivered to Statistics Sweden. Information on VAT is also retrieved from the Customs each six months.

4.A Metadata

Chart 4A. Indicators of output and input data quality of VAT returns – relevance

Indicator	Quality factor	Description
A1	Relevance of population	<p>Definition of the administrative object set. Which rules determine which objects are included? Is this set suitable as statistical population?</p> <p>It is important to note that the VAT-register only (with a few exceptions) includes turnover based on VAT. Services that are not taxable are not included in the register. This means that enterprises in certain sectors are not included in the register, while others may have a turnover which is only partly taxable and thus higher in other administrative sources than the VAT-register.</p> <p>Some legal units report VAT returns as a group and a model is used to distribute the total turnover among the legal units included in this group.</p> <p>The source contains information for only one object set namely the legal unit.</p> <p>A1 = The source contains one kind of relevant populations or important subpopulations.</p>
A2	Relevance of units	<p>Definition of the administrative units. Are these suitable as statistical units?</p> <p>The source contains information on one administrative unit – namely the legal unit. The legal unit is not used as statistical unit but from legal unit both enterprise unit and kind of activity unit can roughly be defined.</p> <p>In the VAT data we obtain information for a group of legal units. A group of legal units can report VAT data just as one legal unit. In the VAT data there is roughly between 1500 -2000 legal units that provide data as a group.</p> <p>A2 = The source contains one unit legal unit – from this most of the enterprise units and kind of activity units can be derived.</p>
A3	Relevant keys	<p>Are there primary keys and foreign keys in the source that are suitable for micro integration within the NSI?</p> <p>There is one important key in this source:</p> <ul style="list-style-type: none"> - Identity number of the legal unit, this is Business Identity Number that is used in the Business register at Statistics Sweden. <p>A3 = The source contains one very relevant key</p>
A4	Relevance of variables	<p>Definitions of the administrative variables. Are these variables suitable as statistical variables?</p> <p>VAT turnover , export of goods (within EU and outside EU) and export of services (within EU and outside EU). The definitions of these variables do not correspond exactly with the definitions in the STS, SBS and the foreign trade statistics. However these differences are in most of the</p>

cases quite small. They are large in activities which are not taxable with VAT.		
A4 = The variables in the source are in most of the cases relevant for statistical purposes.		
A5	Relevance of reference time	Are reference times suitable for statistical usage? What rules for accruing accounting data between months and years are used?
There are three different reference periods. The largest legal units have to provide data monthly, medium sized enterprises quarterly and the smallest one only yearly.		
A5 = Reference time in the source is relevant for yearly statistics. But can be used with rather high quality on a quarterly basis.		
A6	Study domains	Can the units be allocated between relevant study domains?
The VAT-statements can be linked to the Business Register. All classification variables in these registers or in registers that can be linked with these base registers can be used to define study domains.		
A6 = All kinds of study domains are possible for the source.		
A7	Comprehensiveness	Does the source contain a large part of an intended population and many statistically interesting variables? Can many surveys benefit from the source?
The source covers all enterprise that provide VAT turnover. The source contains a few but very important economic variables and is used for many different surveys regarding enterprises.		
A7 = The source is comprehensive it covers all enterprise that provide VAT turnover.		
A8	Updates, delivery time and punctuality	How often and at what time points is the administrative register updated? Time for delivery of the source from register holder to the NSI. Difference in time between delivery and agreed delivery time.
The source is updated weekly, but the data has three different reference periods: month, quarter and year, based upon the projected size of the turnover.		
A8 = The source is updated weekly but the final data for legal units that declare VAT annually is quite late.		
A9	Comparability over time	Extent of changes in the content of the administrative register over time
There are no changes regarding populations and variables that give rise to statistical problems.		
A9 = Comparability over time is good		

The relevance of this source is very high. Data from the source can be used for the business register but can also be used to create frames, stratification for economic surveys but can also be used together with other sources to produce turnover statistics.

4.B Analysis and Data Editing of the Source

Chart 4B. Indicators of output and input data quality of VAT returns – accuracy

Indicator	Quality factor	Description
B1	Primary key	Fraction of units with usable identities. The primary key should have correct format and reasonable values. Enterprise identity number (BIN). All legal units have identity number. B1 = Primary key is usable. All legal units have identity number.
B2	Foreign keys	Fraction of units with usable foreign keys. Foreign keys should have correct format and reasonable values. Link to Business Register – BIN usable B2 = 97,4% of enterprises in VAT-register have usable BIN, the key to the Business Register.
B3	Missing values	Fraction of missing values for the statistically interesting variables. For some VAT returns, there are missing values in some columns. Most of these errors are quite easily corrected by using other data for the relevant month. Overall, less than 1% of all reports for a given period needs correction. B3 = This problem is quite small.
B4	Wrong values	Fraction of erroneous or unreasonable values for the statistically interesting variables. Decimal errors are quite common, i e one post is ten (or a hundred or a thousand) times higher or lower. Another common error is using the wrong column for the VAT-level, for example writing the sum for 25%-level VAT in the 12% column. Since that data is used to calculate turnover, a faulty turnover figure is produced. However, these errors only occur in less than 1% of all Tax reports. B4 = This problem is not so large.
B5	Quality of preliminary data	Fraction of records corrected by the taxpayers. Estimates based on preliminary data are compared with estimates based on final data. Fraction of corrected records by the taxpayers can not be calculated. Preliminary data is overwritten by last known data. The problem is small. B5 = The preliminary data is of high quality.

On the whole accuracy is good.

4.C Integrate the Source with the Base Register

Chart 4C. Indicators on output and input data quality of VAT returns – accuracy

Indicator	Quality factor	Description
C1	Undercoverage in the base register	Fraction of units: There are units that have been active during the reference period but are missing in the Population Register and/or missing or coded as inactive in the Business Register. C1 = The total undercoverage in the Business Register is 10,4%. The undercoverage in the Business Register due to legal units missing is 5,9% The undercoverage in the Business Register due to inactive legal units is 4,5%
C2	Overcoverage in base register	Fraction of units: Enterprises are coded as active in the BR and belong to a category that is covered by the source, but has no activity in the source. C2 = The overcoverage in the Business Register's category of legal units registered for VAT is 7,5%.
C3	Undercoverage in the source	Fraction of units: There are enterprises/units that have been active during the reference period according to the BR but are missing in the source. C3 = No problem.
C4	Overcoverage in the source	Fraction of units: There are units in the source that belong to a category, or seem to belong to a category, that is not statistically relevant. C4 = No problem (see C1)
C5	Can the source improve base register?	Here a more thorough analysis is required depending on the character of the source. The quality improvements should be measured. The source can be used to identify overcoverage in the business register. The source cannot be used to identify undercoverage in the business register. C5 = VAT data can be used to improve business register.

4.D Integrate the Source with Surveys with Similar Variables

Chart 4D. Indicators on input data and production process quality of VAT returns

Indicator	Quality factor	Description	
D1	Is the source good or bad?	a) Compare populations b) Compare units c) Compare variables	Measures production process quality
<p>When turnover is compared with SBS on activity level there can be rather large differences. This is because much of the actual turnover is not VAT taxable.</p> <p>D1 = The source is not so good</p>			
D2	Is the production system good or bad?	a) Compare populations b) Compare units c) Compare variables	Measures production process quality
<p>The coverage for VAT data is very high. However many surveys need the net-turnover not just the VAT based turnover. The current usage of VAT based turnover needs to be replaced by net turnover in some cases.</p> <p>D2 = Not so good for production system use.</p>			
D3	Can the source improve other surveys?	a) Will population be better? b) Will units be better? c) Will variables be better?	Measures production process quality
<p>The source is improving both the external trade statistics and shall possibly be used in the turnover statistics (for medium and small legal units). If the source is compared to SBS, both the VAT and SBS are expected to be improved.</p> <p>D3 = The source can improve other surveys.</p>			
D4	Can the source be combined with other sources?	a) Will population be better? b) Will units be better? c) Will variables be better?	Measures input data quality
<p>The source can be combined with the yearly income declarations for enterprises (SRU) to better cover the economic activity of enterprises in Sweden. There are many enterprises in the VAT register that you can't find in the SRU.</p> <p>D4 = VAT can be combined with other sources.</p>			

The input data quality and production process quality of the VAT data is not good. The source can't be used by itself for turnover statistics. The source is improving both the external trade statistics and shall possibly be used in the turnover statistics (for medium and small legal units).

4.E Quality Report Card for VAT returns

Chart 4E-A. Information from the Administrative Authority – Relevance

Indicator	Quality factor	Description
A1	Relevance of population	The source contains one kind of relevant populations or important subpopulations.
A2	Relevance of units	The source contains one unit legal unit – from this can most of the enterprise units and kind of activity units be derived.
A3	Relevant keys	The source contains one very relevant key
A4	Relevance of variables	The variables in the source are in most of the cases relevant for statistical purposes.
A5	Relevance of reference time	Reference time in the source is relevant for yearly statistics. But can be used with rather high quality on a quarterly basis.
A6	Study domains	All kinds of study domains are possible for the source.
A7	Comprehensiveness	The source is comprehensive it covers all enterprise that provide VAT turnover
A8	Updates, delivery, punctuality	The source is updated weekly but the final data for legal units that declare VAT yearly is quite late.
A9	Comparability over time	Comparability over time is good

Summary: The relevance of this source is very high.

Chart 4E-B. Analysis and Data Editing of the Source – Accuracy

Indicator	Quality factor	Description
B1	Primary key	Primary key is usable. All legal units have identity number.
B2	Foreign keys	97,4% of enterprises in VAT-register have usable BIN, the key to the Business Register.
B3	Missing values	
B4	Wrong values	This problem is not so big.
B5	Quality of preliminary data	The feeling is that the preliminary data is of rather high quality.

Summary: On the whole accuracy is good.

Chart 4E-C. Integrate the Source with the Base Register – Accuracy

Indicator	Quality factor	Description
C1	Undercoverage in base register, fraction of units	10.4%
C2	Overcoverage in BR, fraction of units	7.5%
C3	Undercoverage in the source	No problem
C4	Overcoverage in the source	No problem (see C1)
C5	Can the source improve base register?	VAT data can be used to improve business register.

Summary: On the whole accuracy is good.

Chart 4E-D. Integrate with Surveys with Similar Variables – Production process quality

Indicator	Quality factor	Description
D1	Is the source good or bad?	The source is not so good
D2	Is the production system good or bad?	Not so good for use in the production system
D3	Can the source improve other surveys?	The source can improve other surveys.
D4	Can the source be combined with other sources?	VAT can be combined with other sources.

Summary: The production process quality of the VAT data is not good

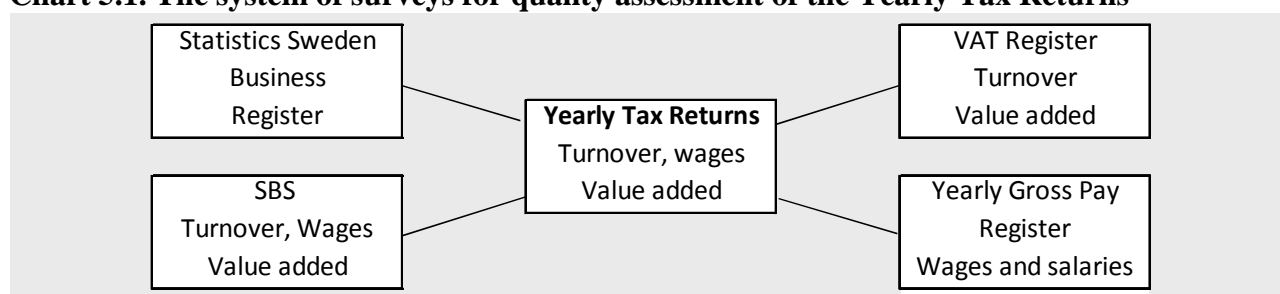
5. Application III – Yearly Tax Returns from Enterprises (SRU)

Yearly Tax Returns should be delivered during May the year after the fiscal year. There are different tax returns for Sole proprietorships, Trading partnerships, Limited partnerships, Limited companies and Economic associations. Yearly Tax Returns consists of three parts, balance sheet, profit and loss statement and tax adjustments. Limited companies give more information and sole traders give less detailed information.

The Yearly Tax Returns is the main source for the Structural Business Statistics survey and the micro-simulation model FRIDA. The Yearly Tax Returns are also used by other surveys.

The indicators in Section 5.C are evaluated by comparisons between the Business Register and Yearly Tax Returns. The indicators in Section 5.D are evaluated by comparisons between Yearly Tax Returns and the VAT Register, the Yearly Gross Pay Register and the Structural Business Statistics survey.

Chart 5.1. The system of surveys for quality assessment of the Yearly Tax Returns



5.A Metadata

The Yearly Tax Returns from enterprises is an administrative source that has been used by Statistics Sweden for many years. The metadata that are the basis of the presentation here consist of Statistics Sweden's experience, tax forms and supporting brochures and contacts with persons at the Swedish Tax Agency.

Chart 5A. Indicators of output and input data quality of Yearly Tax Returns – relevance

Indicator	Quality factor	Description
A1	Relevance of population	Definition of the administrative object set. Which rules determine which objects are included? Is this set suitable as statistical population?
<p>All enterprises must deliver a tax return. There is no informal sector as in some other countries where small enterprises and self-employed in rural or indigenous parts of the economy mustn't have to report tax.</p> <p>A1 = This source contains a relevant population</p>		
A2	Relevance of units	Definition of the administrative units. Are these suitable as statistical units?
<p>The source consists of legal units or legal entities and natural persons. The legal units are administrative objects that can give rise to problems when they are used for statistical purposes. Complex enterprise units consist of many legal units and these enterprises can organise tax reporting so that some units report VAT, another set of units report PAYE and another set of units report Yearly Tax Returns. These problems are discussed in Section 5.D</p> <p>A2 = The source consists of administrative units that are not always suitable for statistical purposes</p>		
A3	Relevant keys	Are there primary keys and foreign keys in the source that are suitable for

micro integration within the NSI?

The identity numbers used in Yearly Tax Returns are the National Identity Numbers given to each legal unit by the Swedish Tax Agency. Sole traders or natural persons use their Personal Identity Numbers given to each person at birth or immigration. These identity numbers are used by Statistics Sweden in all surveys.

A3 = The primary keys in the source are suitable for micro integration within Statistics Sweden. There are foreign keys to partners in Trading and Limited partnerships.

A4 Relevance Definitions of the administrative variables. Are these variables suitable as
 of variables statistical variables?

The source contains variables for profit and loss statements and balance sheets. These variables are important for economic statistics and are necessary for the National Accounts.

A4 = The administrative variables are suitable as statistical variables

A5 Relevance of Are reference times suitable for statistical usage? What rules for accruing
 reference time accounting data between months and years are used?

Many legal units have a financial year that does not correspond to the calendar year. Of the limited companies about 38 % have a broken fiscal year. The turnover of these companies is about 17 % of the total turnover among limited companies. There is information on dates for start and end of the fiscal year and the tax return data can be combined with VAT and Gross Pay data that are monthly. In this way the allocation over time can be estimated.

A5 = For many legal units the fiscal year does not correspond to the calendar year. Estimation methods based on a combination with other administrative sources must be developed to adjust tax return data.

A6 Study domains Can the units be allocated between relevant study domains?

The tax return data can be matched with the Business Register and NACE with information on economic activity and Institutional Sector can be imported into the register with all tax return data. In this way economic statistics for all important domains can be produced.

A6 = The units can be allocated between study domains

A7 Comprehen- Does the source contain a large part of an intended population and many
 siveness statistically interesting variables? Can many surveys benefit from the source?

The tax return source covers the intended population and contains many statistically interesting variables. However, for the National Accounts more detailed information is necessary. For this reason the tax return data are combined with sample survey data containing the detailed information. Estimated ratios from the sample are used to distribute tax return variables into the parts that are needed for the National Accounts.

A7 = The source contains a large part of the intended population and many statistically interesting variables. The National Accounts benefit from the source.

A8 Updates, How often and at what time points is the administrative register updated?
 delivery time Time for delivery of the source from register holder to the NSI. Difference in
 and punctuality time between delivery and agreed delivery time.

A8 = The source is yearly and is delivered during the autumn the year after the reference year.

A9 Comparability Extent of changes in the content of the administrative register over time
 over time

The variable content has been reduced due to the desire to reduce the administrative response burden of the enterprises. But comparability over time is acceptable.

A9 = Only moderate changes over time have taken place.

The relevance of this source, the Yearly Tax Returns from enterprises, is very high. The data source is necessary for the yearly National Accounts.

5.B Analysis and Data Editing of the Source

The data with tax returns consist of data based on different tax forms for different kinds of enterprises or legal units. We use here the names used in Swedish tax administration. Sole traders have the tax form NE, Trading and Limited partnerships have the tax form INK4, Limited companies have the tax form INK2 and Economic associations the tax form INK3. These four kinds of tax return data have been edited separately. Simple subject-matter based automatic editing software has been developed for each kind of tax form. We present here some results from editing of tax forms for limited companies, INK2.

Chart 5B. Indicators of output and input data quality of Yearly Tax Returns – accuracy

B1	Primary key	Fraction of units with usable identities. The primary key should have correct format and reasonable values.
B1 = The primary key has correct format and reasonable values.		
B2	Foreign keys	Fraction of units with usable foreign keys. Foreign keys should have correct format and reasonable values.
B2 = Identity numbers of partners for Trading and Limited partnership have correct format		
B3	Missing values	Fraction of missing values for the statistically interesting variables.
Information on the fiscal year is missing in about 24 % of the cases. This is interpreted as that fiscal and calendar years are identical. When cells regarding economic variables are empty this is interpreted as zero values.		
B3 = Missing values is not an important issue.		
B4	Wrong values	Fraction of wrong or unreasonable values for the statistically interesting variables.
The statistically interesting variables in the balance sheet and the profit and loss statement have a small number of errors. But some of these errors can be very large. It is easy to find and correct these errors and after these corrections the tax return data have acceptable quality.		
B4 = Small fraction of very large errors that are easy to find and correct.		
B5	Quality of preliminary data	Fraction of records corrected by the taxpayers. Estimates based on preliminary data are compared with estimates based on final data.
For the tax form INK2 for limited companies, the fraction of data that was corrected of the editing software was 9.4 % using the preliminary data. In the final data, this fraction had dropped to 4.7 %. For the tax form NE for sole traders, the fraction of corrected data was 29% using the preliminary data. In the final data, this fraction had dropped to 16%.		
B5 = Preliminary data is delivered to Statistics Sweden during August; the final data with better quality is available during December. December is in time for the users of Tax Return data.		

The accuracy of the Yearly Tax Returns is acceptable if the data are edited to remove large technical errors and sign errors. Timeliness of the final version of this data is sufficient.

5.C Integrate the Source with the Base Register – the Business Register

The object set with Yearly Tax Returns has been compared with the November Frame of the Business Register. This November frame is currently used for the Structural Business Statistics survey.

Chart 5C. Indicators on output and input data quality of Tax Returns – accuracy

Indicator	Quality factor	Description
C1	Undercoverage in the base register	Fraction of units: There are units that have been active during the reference period but are missing in the Population Register and/or missing or coded as inactive in the Business Register.
<p>The object set with all Yearly Tax Returns was compared with the November Frame of the Business Register. 27.4 % of the legal units that had delivered a Yearly Tax Form were not in the November frame. 14.8 % were coded as inactive and 12.6 % were missing completely.</p> <p>C1 = The total undercoverage in the Business Register is 27.4 %. The undercoverage in the Business Register due to inactive legal units is 14.8 % and the undercoverage in the Business Register due to missing legal units is 12.6 %</p>		
C2	Overcoverage in base register	Fraction of units: Enterprises are coded as active in the BR and belong to a category that is covered by the source, but has no activity in the source.
<p>C2 = The overcoverage in the BR is 15.4 %, i.e. were “active” in the BR but did not report a Tax Return.</p>		
C3	Undercoverage in the source	Fraction of units: There are enterprises/units that have been active during the reference period according to the BR but are missing in the source.
<p>C3 = No sign of undercoverage in the object set of all Yearly Tax Returns.</p>		
C4	Overcoverage in the source	Fraction of units: There are units in the source that belong to a category, or seem to belong to a category, that is not statistically relevant.
<p>C4 = Some Yearly Tax Returns can be ostensible transactions with the purpose to escape from taxation</p>		
C5	Can the source improve base register?	Here a more thorough analysis is required depending on the character of the source. The quality improvements should be measured.
<p>The Yearly Tax Returns can be used to improve coverage of the Business Register. Small new enterprises can appear for the first time in this source. Also, the turnover in the Tax Returns have better coverage than the turnover in the VAT returns, which are currently used in the BR.</p> <p>C5 = Yes</p>		

Coverage errors in the Business Register were found after comparisons with the Yearly Tax Returns.

5.D Integrate the Source with Surveys with Similar Variables

To evaluate quality indicator D1 we have compared the Yearly Tax Returns with two administrative sources containing similar variables. Turnover is included both in Yearly Tax Return Register and in the VAT Register. Yearly gross pay is included both in Yearly Tax Returns and the Yearly PAYE Register. To evaluate quality indicator D2 the Structural Business Statistics survey (SBS) has been compared with the Yearly Tax Return Register, the VAT Register and Yearly PAYE Register. These comparisons are also that basis for the conclusions regarding indicators D3 and D4.

Chart 5D. Indicators on input data and production process quality of Yearly Tax Returns

Indicator	Quality factor	Description	
D1	Is the source good or bad?	a) Compare populations b) Compare units c) Compare variables	Measures production process quality
<p>a) Populations: The administrative object set for the Yearly Tax Returns is difficult to use as statistical population in combination with the NACE codes in the Business Register due to the fact that large enterprises report different economic variables using different legal units</p> <p>b) Units: The administrative objects in the Yearly Tax Returns Register are difficult to use in combination with other sources. Complex enterprise units must be created to avoid errors and inconsistencies.</p> <p>c) Variables: The variables in the Yearly Tax Returns Register are useful, no indication of bad quality have been found when turnover and gross pay were compared with other sources.</p> <p>D1 = The administrative units in the Yearly Tax Return Register make that this source is difficult to use for statistical purposes. Without appropriate methods estimates by industry will have errors.</p>			
D2	Is the production system good or bad?	a) Compare populations b) Compare units c) Compare variables	Measures production process quality
<p>a) Populations: The present production system for the SBS-survey should be improved. The present population is not as good as it could be.</p> <p>D2 = The population for the SBS-survey has undercoverage errors and overcoverage errors.</p>			
D3	Can the source improve other surveys?	a) Will population be better? b) Will units be better? c) Will variables be better?	Measures production process quality
<p>a) Populations: The population of the SBS-survey can be improved if the whole Yearly Tax Return register, the VAT-register and the PAYE-register are used.</p> <p>b) Units: Much more efforts should be spent on creating complex enterprise units, otherwise estimates by industry will have poor quality.</p> <p>c) Variables: The VAT-register and the PAYE-register can be used to estimate calendar year variable values for enterprises with broken fiscal year.</p> <p>D3 = The SBS-survey can be improved: Better population, better statistical units and better variables can be created if Yearly Tax Returns, the VAT-register and the PAYE-registers are used in the right way.</p>			
D4	Can the source be combined with other sources?	a) Will population be better? b) Will units be better? c) Will variables be better?	Measures input data quality

a) Populations: A calendar year version of the Business Register can be created with the administrative sources described here. Also data from the Customs and some other administrative sources can be included. This register can be created during November the year after the reference year. This register can be the platform for the data that should be delivered to the yearly National Accounts that need the data at this time.

b+c) Units and variables: If similar variables from different sources are compared it will be possible to see which legal units that must be combined into complex enterprise units. In this way consistency between variables will be attained and estimates by industry will be consistent for all variables in the register.

D4 = The Yearly Tax Returns should be combined with other relevant administrative sources and a calendar year register with all enterprises active during some part of the calendar year should be created.

The Yearly Tax Returns is a very complicated set of data that requires that good statistical methods are developed. The population needs special methods and the units must be improved.

5.E Quality Report Card for Yearly Tax Returns

Chart 5E-A. Information from the Administrative Authority – Relevance

Indicator	Quality factor	Description
A1	Relevance of population	This source contains a relevant population
A2	Relevance of units	The administrative units (legal units) are not always suitable for statistical purposes
A3	Relevant keys	The primary keys are suitable for micro integration. There are suitable foreign keys to partners in Trading and Limited partnerships.
A4	Relevance of variables	The administrative variables are suitable as statistical variables
A5	Relevance of reference time	For many legal units the fiscal year does not correspond to the calendar year. Estimation methods based on a combination with other administrative sources must be developed to adjust tax return data.
A6	Study domains	The units can be allocated between study domains
A7	Comprehensiveness	The source contains a large part of the intended population and many statistically interesting variables. The National Accounts benefit from the source.
A8	Updates, delivery, punctuality	The source is yearly and is delivered during the autumn the year after the reference year.
A9	Comparability over time	Only moderate changes over time have taken place.

Summary: The relevance of this source, the Yearly Tax Returns from enterprises, is very high. The data source is necessary for the yearly National Accounts.

Chart 5E-B. Analysis and Data Editing of the Source – Accuracy

Indicator	Quality factor	Description
B1	Primary key	The primary key has correct format and reasonable values.
B2	Foreign keys	Identity numbers of partners for Trading and Limited partnership have correct format
B3	Missing values	Missing values is not an important issue.
B4	Wrong values	Small fraction of very large errors that are easy to find and correct.
B5	Quality of preliminary data	Preliminary data is delivered during August; the final data with better quality is available during December.

Summary: The accuracy of the Yearly Tax Returns is acceptable if the data are edited to remove large technical errors and sign errors.

Chart 5E-C. Integrate the Source with the Base Register – Accuracy

Indicator	Quality factor	Description
C1	Undercoverage in base register, fraction of units	The total undercoverage in the Business Register is 27.4 %.
C2	Overcoverage in BR, fraction of units	The overcoverage in the BR is 15.4 %, i.e. were “active” in the BR but did not report a Tax Return.
C3	Undercoverage in the source	No sign
C4	Overcoverage in the source	Some Tax Returns can be ostensible transactions
C5	Can the source improve base register?	The coverage and turnover measure can be improved

Summary: Coverage errors in the Business Register were found after comparisons with the Yearly Tax Returns.

Chart 5E-D. Integrate with Surveys with Similar Variables – Production process quality

Indicator	Quality factor	Description
D1	Is the source good or bad?	The administrative units make that this source is difficult to use for statistical purposes. Without appropriate methods estimates by industry will have errors.
D2	Is the production system good or bad?	The population for the SBS-survey has undercoverage errors and overcoverage errors.
D3	Can the source improve other surveys?	The SBS-survey can be improved: Better population, better statistical units and better variables can be created
D4	Can the source be combined with other sources?	The Yearly Tax Returns should be combined with other sources and a calendar year register should be created.

Summary: The Yearly Tax Returns is a very complicated set of data that requires the development of good statistical methods. The population needs special methods and the units must be improved.

6. Discussion

This report includes three examples of applications of the register quality indicators suggested by Laitila et al. (2011, 2012). The indicators are grouped in sets which provides with a working procedure involving the sequence of evaluating metadata, accuracy, integration with a base register, and integration with other surveys. The definitions of the indicators are simple and direct and suggested with the view of implementing a register within statistical register systems as proposed by Wallgren and Wallgren (2007).

It is natural to first consider the contents of a register and then evaluate the accuracy of the contents. For integration it is of interest to know if the register can be incorporated into the system by relating it to a base register. Performing these three first steps of the evaluation process does not only cast light on the new source evaluated. The experiences from the examples are that the evaluation process also provides with insights on the registers and other data sources used in the evaluation process. This is perhaps one of the more important results in the applications reported and an issue for future projects. Keeping a register system it should continuously be evaluated with respect to consistency with relevant external information. Here any kind of relevant information is of interest and not only administrative registers potentially useful for statistics production.

The fourth set of indicators addresses the issue of how a register can be utilized for improving surveys conducted at an NSI. Again the experience of the applications is that this step provides with both information on the quality of the register evaluated and the data sources used for the evaluation. Another expected experience is that this evaluation is a methodologically demanding task. This step will also involve subjective judgments since the considered potential usage of the new register depends on the experiences of those performing the evaluation.

References

- Daas, P.J.H, Arends-Tóth, J., Schouten, B. and L. Kuijvenhoven (2008). Proposal for a quality framework for the evaluation of administrative and survey data, Paper for the workshop “Combination of surveys and administrative data”, May 2008, Vienna, Austria.
- Eurostat (2003). Quality Assessment of Administrative Data for Statistical Purposes; Working group “Assessment of Quality in Statistics”, Luxembourg, 2-3 October, 2003. Web publication, Eurostat.
- Kott, P.S. and T. Chang (2010). Using calibration weighting to adjust for nonignorable unit nonresponse, *Journal of the American Statistical Association*, **105:491**, 1265-1275.
- Laitila, T, Wallgren A. and B. Wallgren (2011). Quality Assessment of Administrative Data, Research and Development – Methodology reports from Statistics Sweden, 2011:2.
- Laitila, T. Wallgren A. and B. Wallgren (2012). Assessment of Administrative Data Source Quality, Project Deliverable 4.3 – part 2, FP7 - BLUE-ETS, European Commission.
- Lehtonen, R. and A. Veijanen (1999). Use of Register Data to Improve the Estimation in a Sample Survey, In Aho, J. (Ed.) *Statistics, Registries, and Science – Experiences from Finland*, Statistics Finland.
- Särndal, C.-E., Swensson, B. and J. Wretman (1992). *Model Assisted Survey Sampling*, Springer, New York.
- Särndal, C.-E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.
- Wallgren, A. and B. Wallgren (2007). *Register-based Statistics – Administrative Data for Statistical Purposes*. John Wiley & Sons Ltd, Chichester, England.

Appendix D: Manual for determining the input quality of administrative data sources with the ‘dataquality’ package including a Quality Report Card.



Funded under Socio-economic Sciences & Humanities

EUROPEAN COMMISSION
RESEARCH DIRECTORATE-GENERAL



BLUE-Enterprise and Trade Statistics
BLUE-ETS

SP1-Cooperation-Collaborative Project
Small or medium-scale focused research project

FP7-SSH-2009-A
Grant Agreement Number 244767
SSH-CT-2010-244767

Manual

Title: Manual for determining the input quality of administrative data sources with the ‘dataquality’ package including a Quality Report Card.

Authors: Piet Daas, Saskia Ossen and Martijn Tennekes (CBS)

Version: 1.1

DATE 25-01-2013

Manual

Manual for determining the input quality of administrative data sources with the ‘dataquality’ package including a Quality Report Card.

Summary:

In WP4 of the BLUE-ETS project measurement methods were developed for determining the quality of administrative data when used as an input source for the statistical process of National Statistical Institutes. These measurement methods form the basis for the script included in the R-package ‘dataquality’. How to use the implemented measurement methods is described in this manual and illustrated with examples. The manual also contains the most recent version of the Quality Report Card for Administrative data; a card that is used to provide an overview of the findings of the evaluation and guides the user on the decisions to take.

Index

1. Introduction.....	153
2. Implementation of measurement methods.....	153
2.1. Description of synthetic data.....	153
2.2. Technical checks.....	154
2.2.1. Readability.....	154
2.2.2. File declaration compliance.....	155
2.2.3. Convertability.....	158
2.3. Accuracy.....	158
2.3.1. Authenticity.....	159
2.3.2. Inconsistent objects.....	161
2.3.3. Dubious objects.....	163
2.3.4. Measurement error.....	163
2.3.5. Inconsistent values.....	163
2.3.6. Dubious values.....	165
2.4. Completeness.....	165
2.4.1. Undercoverage.....	166
2.4.2. Overcoverage.....	166
2.4.3. Selectivity.....	167
2.4.4. Redundancy and missing values.....	168
2.4.5. Imputed values.....	170
2.5. Time-related.....	170
2.5.1. Timeliness, Punctuality, Overall time lag, Delay.....	171
2.5.2. Dynamics of objects.....	172
2.5.3. Stability of variables.....	173
2.6. Integrability.....	174
2.6.1. Comparability of objects.....	176
2.6.2. Alignment of objects.....	177
2.6.3. Linking variable.....	177
2.6.4. Comparability of variables.....	177
3. The Quality Report Card.....	177
References.....	178
Quality Report Card for Administrative data.....	180
1. Technical checks.....	182
2. Time-related dimension.....	186
3. Completeness.....	185
4. Accuracy.....	184
5. Integrability.....	Fout! Bladwijzer niet gedefinieerd.
6. General findings.....	187

1. Introduction

National Statistical Institutes (NSI's) that want to increase the use of administrative sources (i.e. registers) for statistical purposes need methods to evaluate the quality of those sources from a statistical point of view. This topic was studied in the European funded BLUE Enterprise and Trade Statistics project (BLUE-ETS, 2011) and started by identifying the quality 'components' that determine the input quality of administrative data sources (Daas *et al.*, 2011a). Next, measurement or estimation methods were proposed for the quality indicators identified (Daas *et al.*, 2011b). By implementing them in software, in a R-package called 'dataquality', it was assured that of all the (theoretically) developed methods those remained that could actually be used. How these methods remaining should be used is described in this report and illustrated with examples. In addition, a Quality Report Card for Administrative data (QRCA) is included. The report card is used to provide an overview of all evaluation results obtained and guides the user in the subsequent steps to take.

2. Implementation of measurement methods

This chapter discusses for every measurement or estimation method whether it can be implemented, whether it is already implemented and how. The methods are discussed per dimension discerned for the input quality of administrative data (Daas *et al.*, 2011a). These are:

- *Technical checks*, technical usability of the file and data in the file (paragraph 2.2)
- *Accuracy*, the extent to which data are correct, reliable, and certified (paragraph 2.3)
- *Completeness*, degree to which a data source includes data describing the corresponding set of real-world objects and variables (paragraph 2.4)
- *Time-related*, indicators that are time and/or stability related (paragraph 2.5)
- *Integrability*, extent to which the data source is capable of undergoing integration or of being integrated (paragraph 2.6)

The methods implemented in the 'dataquality' R-package are illustrated by applying them to synthetic data. The dataquality package will be made publically available by publishing it on the Comprehensive R Archive Network in due time. The most recent version can also be obtained by contacting the authors of this manual.

2.1. Description of synthetic data

To be able to fully illustrate the methods implemented three synthetic data files are used:

- *Sam*: the population register of the fictive island Samplonia. This file contains the following variables:
 - Sort (datatype: numeric)
 - Id (datatype: categorical)
 - Province (datatype: categorical)
 - Sex (datatype: categorical)
 - Age (datatype: numeric)
 - Employed (datatype: categorical)
 - Income (datatype: numeric)
 - Education (datatype: categorical)
 - Blood_group (datatype: categorical)
 - Household (datatype: numeric)
- *Reg_tI*: blood group register of Samplonia at time instant t_I . This register contains the following variables:
 - ID_code (datatype: categorical)
 - Sex (datatype: categorical)
 - Age (datatype: numeric)

- Blood_group (datatype: categorical)
- *Reg_t2*: blood group register of Samplonia at time instant t_2 . This register contains the following variables:
 - ID_code (datatype: categorical)
 - Sex (datatype: categorical)
 - Age (datatype: numeric)
 - Blood_group (datatype: categorical)

More details of the data will be shown in the following paragraph, in which a first data exploration is performed as part of the technical checks dimension.

2.2. Technical checks










Table 1 gives an overview of the measurement methods in the Technical checks dimension. A  in the “Can be implemented” column means that the method can be implemented in R. A  in the “Has been implemented” column means that the method has been implemented in the first version of the R package. A  implies that no separate function has been developed for the method, but that the method can be evaluated using functions implemented for other measurement methods. Cells that are not marked can not be implemented.

Table 1. Extent to which the indicators in the Technical checks dimension can be implemented

Dimension indicators	Methods	Can be implemented?	Has been implemented?
1.1 Readability	% of deliveries (or files) of the total deliveries with an unknown extension, that are corrupted, or cannot be opened		
	% of the total file which is unreadable (in size (MB/GB) or number of readable file records)		
1.2 File declaration compliance	% of variables in the current delivery that differ from the metadata lay-out agreed upon in: <ul style="list-style-type: none"> i) formats and names ii) variable and attribute content iii) categories defined for categorical variables iv) ranges for numerical variables (if applicable, e.g. for age: 0-120) 		 describe visualize tableplot
1.3 Convertability	% of objects with decoding errors or corrupted data		

2.2.1. Readability

To be able to apply the measurement methods mentioned in this report to a file, it must be possible to open the file in the software package used. Opening the file is thus the starting point for all other measurement methods. R can be used to read data of any kind of format. However, other software tools may be more suitable to access the data in the original format, report corrupted files, and eventually convert the data to a convenient format. If R is used to read and convert the data, we recommend the packages LaF to process large ASCII files (Van der Laan, 2012), and RSQLite (James, 2011) and RODBC to connect to databases (Ripley and Lapsley, 2012).

2.2.2. File declaration compliance

File declaration compliance can first be checked by applying the function “describe”. Examples of the output of this function for the aforementioned synthetic input files are provided in Figure 1, Figure 2, and Figure 3 for respectively *sam*, *reg_t1*, and *reg_t2*.

For the categorical variables (Id, Province, Sex, Employed, Education, Blood_group, ID_code) a frequency distribution is shown. When there are too many categories to show the complete frequency distribution, the lowest and highest values are shown instead.

```

Console C:/Users/whcg/Desktop/Saskia/
> describe(sam)
...
10 variables      1005 observations
-----
sort
  n missing  unique   Mean   .05   .10   .25   .50   .75   .90   .95
1005         0   1005   503  51.2 101.4 252.0 503.0 754.0 904.6 954.8
lowest :    1    2    3    4    5, highest: 1001 1002 1003 1004 1005
-----
id
  n missing  unique
1005         0   1000
lowest : s1    s10   s100 s1000 s101 , highest: s995 s996 s997 s998 s999
-----
province
  n missing  unique
1005         0     7
      Akkerwinde Grasmalen Nieuwekans Lommerdal Smeulde stapelrade vuilpanne
Frequency      145      95      55      62      245      148      255
%              14       9       5       6       24       15       25
-----
sex
  n missing  unique
1005         0     2
male (513, 51%), female (492, 49%)
-----
age
  n missing  unique   Mean   .05   .10   .25   .50   .75   .90   .95
1005         0     94  34.39    2    5    14    31    52    70    78
lowest :  0  1  2  3  4, highest: 89 90 91 93 97
-----
employed
  n missing  unique
997         8     2
yes (334, 34%), no (663, 66%)
-----
income
  n missing  unique   Mean   .05   .10   .25   .50   .75   .90   .95
997         8    310  419.5    0    0    0    0    520  1602  2248
lowest :    0  101  102  108  109, highest: 4177 4309 4464 4482 4497
-----
education
  n missing  unique
1005         0     6
no formal education (231, 23%), primary (134, 13%), lower secondary (336, 33%), upper secondary (44, 4%)
bachelor (214, 21%), master (46, 5%)
-----
blood_group
  n missing  unique
1005         0     2
A (755, 75%), o (250, 25%)
-----
household
  n missing  unique   Mean   .05   .10   .25   .50   .75   .90   .95
1005         0    324  168.8    16   30   79   167   255   314   331
lowest :    1    2    3    4    5, highest: 346 347 348 349 350
-----
> |

```

Figure 1. Example of the output of the function “describe” that can be used to check the file declaration compliance (applied to file *sam*)

```

> describe(reg_t1)
...
 4 variables      402 observations
-----
id_code
  n missing  unique
 402         0    401

lowest : A111 s1   s101 s105 s106, highest: s93  s95  s96  s97  s98
-----
sex
  n missing  unique
 399         3     2

male (208, 52%), female (191, 48%)
-----
age
  n missing  unique  Mean   .05   .10   .25   .50   .75   .90   .95
 397         5     78  43.72  16.0  17.0  26.0  41.0  59.0  74.0  80.4

lowest : 12 13 14 15 16, highest: 89 90 91 93 97
-----
blood_group
  n missing  unique
 402         0     2

A (201, 50%), O (201, 50%)
-----
> |

```

Figure 2. Example of the output of the function “describe” that can be used to check the file declaration compliance (applied to file *reg_t1*)

```

> describe(reg_t2)
...
 4 variables      403 observations
-----
id_code
  n missing  unique
 403         0    402

lowest : s1   s1001 s101  s105  s106 , highest: s95  s96  s97  s98  s989
-----
sex
  n missing  unique
 400         3     2

male (209, 52%), female (191, 48%)
-----
age
  n missing  unique  Mean   .05   .10   .25   .50   .75   .90   .95
 398         5     80  43.33  15.00  17.00  26.00  41.00  58.75  74.00  80.30

lowest : 0 3 12 13 14, highest: 89 90 91 93 97
-----
blood_group
  n missing  unique
 403         0     2

A (203, 50%), O (200, 50%)
-----
> |

```

Figure 3. Example of the output of the function “describe” that can be used to check the file declaration compliance (applied to file *reg_t2*)

The short summary already gives a first impression of the quality of the data. Several errors, like unexpected values or a high number of missing values can already be detected in this way. The short variable summaries also contain the number of missing values and the number of unique elements. For numerical variables several descriptive statistics like the mean, the sum, and percentiles are indicated next to the number of missings, and the number of unique elements.

These first data explorations in the technical checks dimension can be further extended by visualizing the data with the functions “visualise”, and “tableplot”. The functions are in Figure 4, and Figure 5 applied to the synthetic input file *sam*.

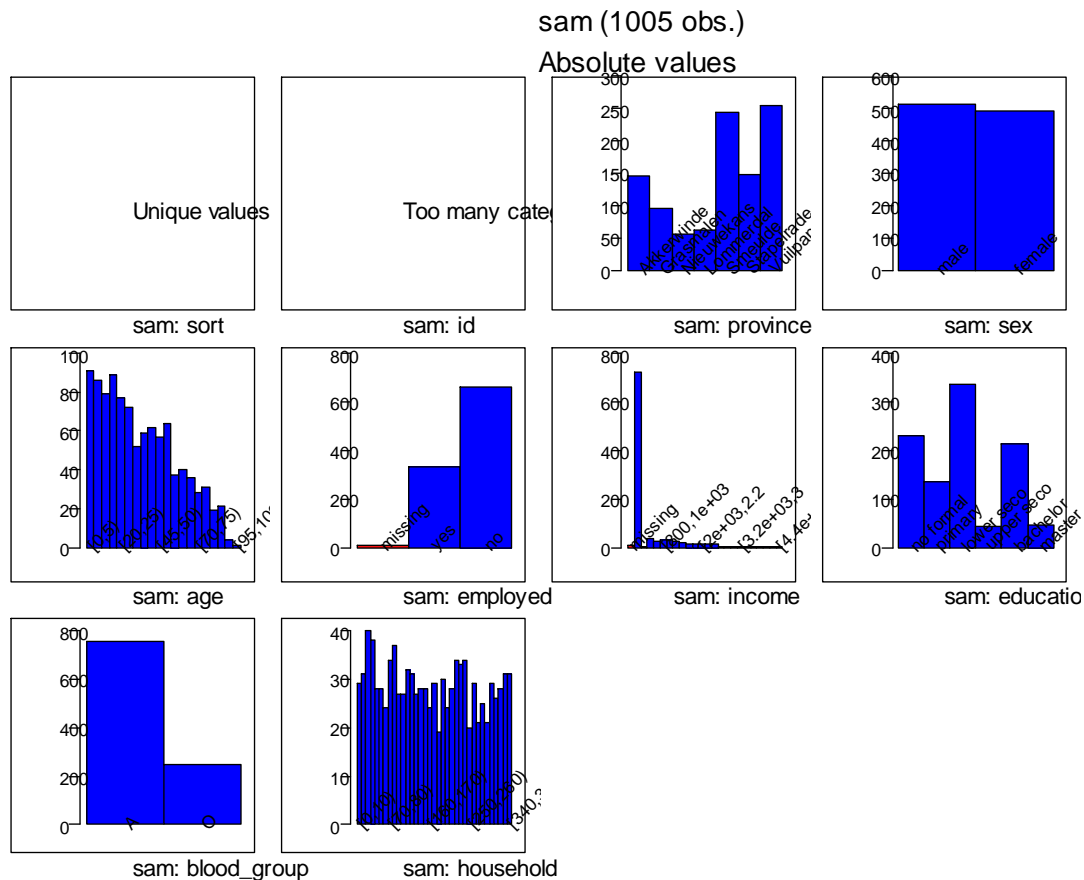


Figure 4. Example of the output of the function “visualise” that can be used to check the file declaration compliance (applied to file *sam*)

Figure 4 shows for every categorical variable a barplot and for every numerical variable a histogram. This gives a quick impression of the values of the different variables in the file.

Figure 5 shows a so-called tableplot (Tennekes *et al.*, 2011, Tennekes, and de Jonge, 2012, Tennekes *et al.*, 2013). A tableplot is a visualization of (large) multivariate datasets. Each column represents a variable and each row bin is an aggregate of a certain number of records. For numeric variables, a bar chart of the mean values is depicted. For categorical variables, a stacked bar chart is depicted of the proportions of categories. Missing values are taken into account. Such a plot can, among others, be used to get an impression of the relations between variables and the presence of (selectively) missing values. The data are in this example sorted on the variable Sort. Since this variable is just the record number no clear relations between it and the other variables can be seen.

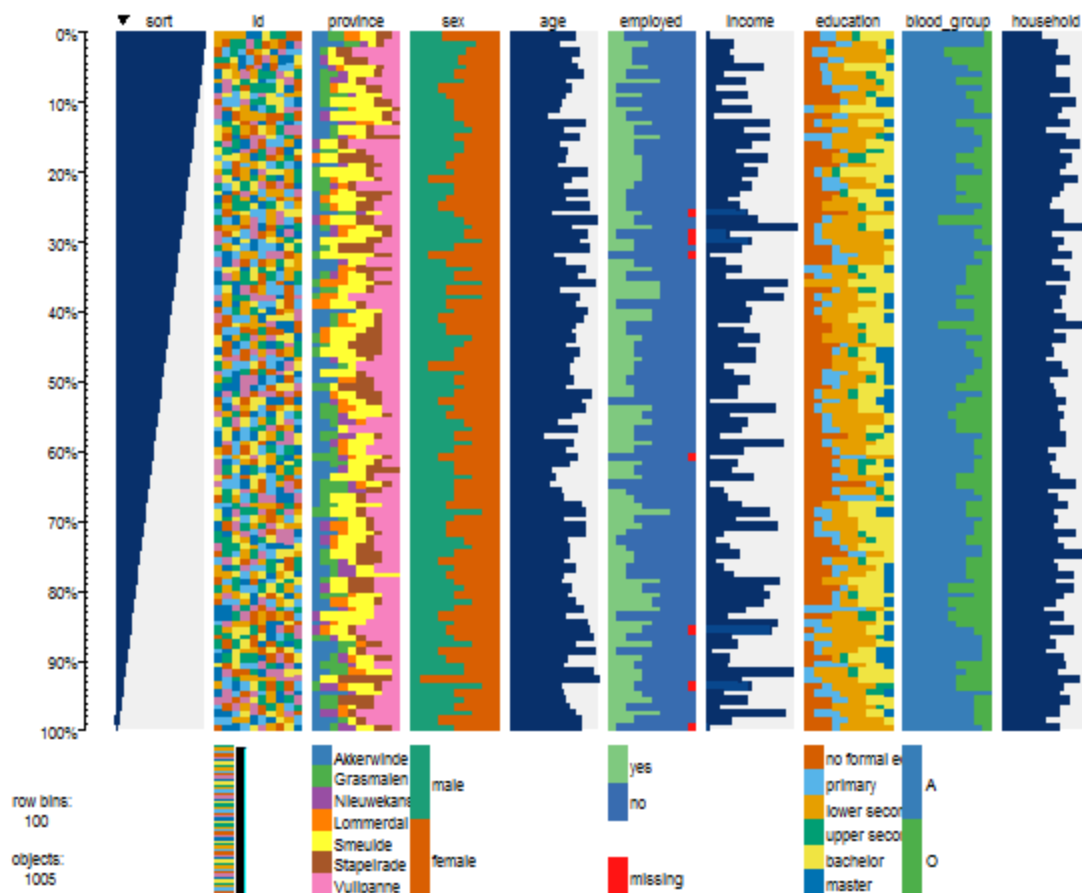


Figure 5. Example of the output of the function “tableplot” that can be used to check the file declaration compliance (applied to file *sam*)

2.2.3. Convertability

The extent to which files can be converted correctly can be determined using the same functions as the ones used for testing for “readability”, and “file declaration compliance” (section 2.2.2). The results before and after conversion can simply be compared, when no other data operations are performed in between any differences are caused by the conversion.

2.3. Accuracy

Table 2 gives an overview of the measurement methods in the Accuracy dimension, and the extent to which they can be implemented.

Table 2. Extent to which the indicators in the Accuracy dimension can be implemented

Dimension indicators	Methods	Can be implemented?	Has been implemented?
2.1 Authenticity	% of objects with a syntactically incorrect identification key	✓	✓ checkFormat
	% of objects for which the data source contains information contradictory to information in a reference list for those objects	✓	✓ changed_value compare
	Contact the data source holder for their % of non-authentic objects in the source		

2.2 Inconsistent objects	% of objects involved in non-logical relations with other (aggregates of) objects		relation
2.3 Dubious objects	% of objects involved in implausible but not necessarily incorrect relations with other (aggregates of) objects		relation
2.4 Measurement error	<p>Only applicable when values not containing measurement errors are marked.</p> <p>% of unmarked values in the data source for each variable</p> <hr/> <p>Contact the data source holder and ask the following data quality management questions:</p> <ul style="list-style-type: none"> - Do they apply any design to the data collection process (if possible)? - Do they use a process for checking values during the reporting phase? - Do they use a benchmark for some variables? - Do they use a checking process for data entry? - Do they use any checks for correcting data during the processing or data maintenance? 		
2.5 Inconsistent values	% of objects of which combinations of values for variables are involved in non-logical relations		Via editrules package
2.6 Dubious values	% of objects with combinations of values for variables that are involved in implausible but not necessarily incorrect relations		Via editrules package

The measurement methods are now discussed in more detail.

2.3.1. Authenticity

The “% of objects with a syntactically incorrect identification key” can be determined after specifying how the key should look like using the function “checkFormat”. Figure 6 shows examples in which the formats of the variables *Id* (*sam*), and *Id_code* (*reg_t1*, *reg_t1*) are checked. The values for these variables should start with a “S” followed by a number. The first part of the example shows that all values for ID in the file *sam* fulfill the requirement. In the file *reg_t1* there is a problem with one value: A111.

```

> #check if the values in the id column of sam precede with an "s", than followed by a
number:
> checkFormat(sam, "id", pattern="^S[0-9]+$")
-----variable id
Number of objects: 1005
Number of missing values: 0 (0%)
Pattern: ^S[0-9]+$
Number of valid values: 1005 (100%)
Number of invalid values: 0 (0%)
Invalid values:
>
> #the same for reg_t1:
> checkFormat(reg_t1, "id_code", pattern="^S[0-9]+$")
-----variable id_code
Number of objects: 402
Number of missing values: 0 (0%)
Pattern: ^S[0-9]+$
Number of valid values: 401 (99.75%)
Number of invalid values: 1 (0.25%)
Invalid values: All

```

Figure 6. Example of the output of the function “checkFormat” that can be used to check whether values for a variable are of a correct format

The implementation of the measurement method “% of objects for which the data source contains information contradictive to information in a reference list for those objects” is illustrated in Figure 7 and Figure 8. In these examples we assume that the file *sam* is the reference list to which the file *reg_t1* is compared. In Figure 6 the function “changed_value” is used to determine the number of different values between the variables Sex and Age in the two different date sources. To be able to do so the sources are linked to each other (Id in *sam* to Id_code in *reg_t1*), and for those records that can be linked the values are compared between the sources.

The same comparison can be done visually using the function “compare”. For the categorical variable Sex this is illustrated in the fluctuation plot shown in the left part of Figure 8, while the right part of this figure displays the result for the numerical variable Age (created using the VIM package of Templ *et al.*, 2012). For this numerical variable, a scatter plot with box plots is shown. The scatterplot shows the values for the file *reg_t1* on the x-axis, and the values for the file *sam* on the y-axis. Scatters on the x=y line have the same value in both sources. The boxplots summarize the distributions of the values in both sources. The red boxplot shows the distribution of values for which a value is missing in the file *reg_t1*.

```

> changed_value(reg_t1, sam, source_t1.keys="id_code", source_t2.keys="id", variable="age")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 5 duplicated records. These records are omitted.
Keys used: id_code ( id )
variable compared: age
Number of records changed: 5 (of 400 aligned records)
Fraction of records changed: 0.0125 , ( 1.25 %)
> changed_value(reg_t1, sam, source_t1.keys="id_code", source_t2.keys="id", variable="sex")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 5 duplicated records. These records are omitted.
Keys used: id_code ( id )
variable compared: sex
Number of records changed: 3 (of 400 aligned records)
Fraction of records changed: 0.0075 , ( 0.75 %)
> |

```

Figure 7. Example of the output of the function “changed_value” that can be used to check whether values in a file are contradictive to values in another file

reg_t1 compared to sam (joined by id_code)
400 common observations

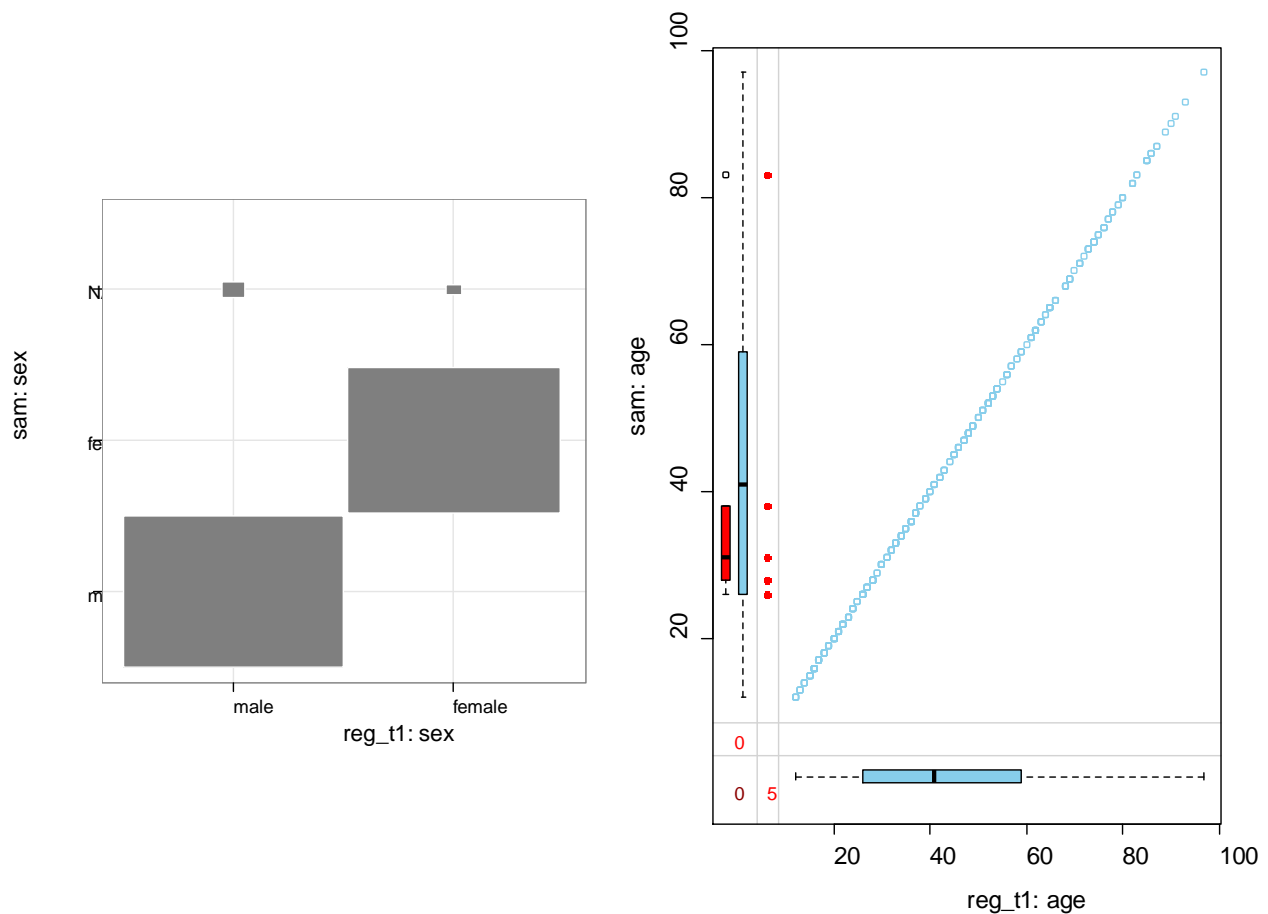


Figure 8. Example of the output of the function “compare” that can be used to check whether values in a file are contradictory to values in another file (variables “sex”, and “age” are considered)

From the figure it can be concluded, that there are no records for which the value for Sex is inconsistent between both files. Differences are caused by a missing value (NA) in either of the files.

2.3.2. Inconsistent objects

The measurement method “% of objects involved in non-logical relations with other (aggregates of) objects” consists of two parts, i.e. it is possible to check for (1) 1:n relations, and (2) m:1 relations. These relations can in table format, and in graphical format be explored using the function “relation”.

Checking for 1:n relations means in this example that we consider whether the file contains persons belonging to more than 1 household. Checking for m:1 relations means that we consider how many persons belong to every household. Figure 9, Figure 10, and Figure 11 show the results of applying the function “relation”.

Suppose that it is not possible that a person belongs to more than 1 household. The lower part of Figure 9 immediately shows that this rule is violated five times. Suppose further that there is a rule stating that households can not consist of 7 people or more. From the upper part of Figure 9 it is clear that this rule is not satisfied by 11 households (7 + 4).

These findings are also visualized by the function “relation”. Figure 10 shows the Id’s of all persons belonging to more than one household. From this figure it is immediately clear that person S688 belongs to household 42 as well as to household 248 (bottom right corner).

Too large households (7 or more persons) are presented in Figure 11. The example shows that household 289 contains 7 persons who have the following Id’s: S649, S774, S511, S580, S597, S285, S419.

```
> relation(sam, "id", "household", graph.lowerbound.numberv2perv1=2, draw=TRUE) # plot persons who belong to
multiple households
Number of units id: 1000
Number of units household: 324
-----
Number of units id per unit household:
      n missing  unique   Mean
324      0         8     3.102

      1 2 3 4 5 6 7 8
Frequency 56 83 59 63 39 13 7 4
%         17 26 18 19 12  4 2 1
-----
Number of units household per unit id:
      n missing  unique   Mean
1000      0         2     1.005

1 (995, 100%), 2 (5, 0%)
```

Figure 9. (upper part) Number of persons per household, and (lower part) numbers of households per person

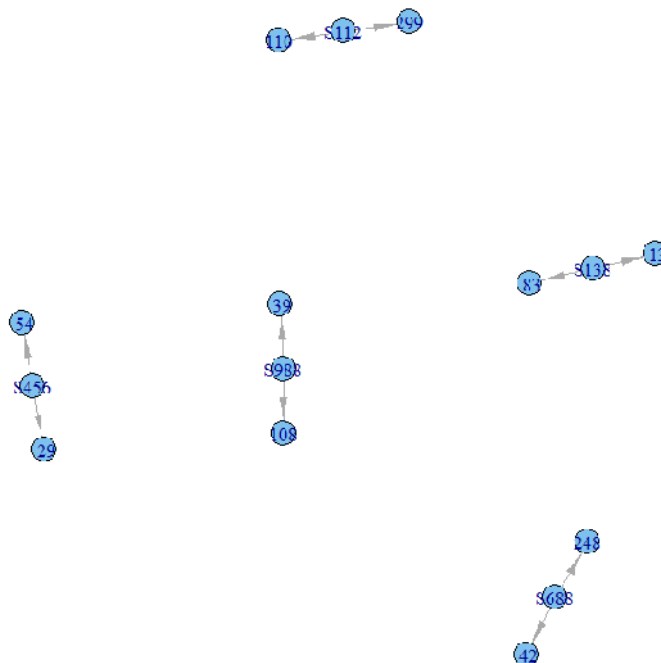


Figure 10. Graphical presentation of persons belonging to more than 1 household

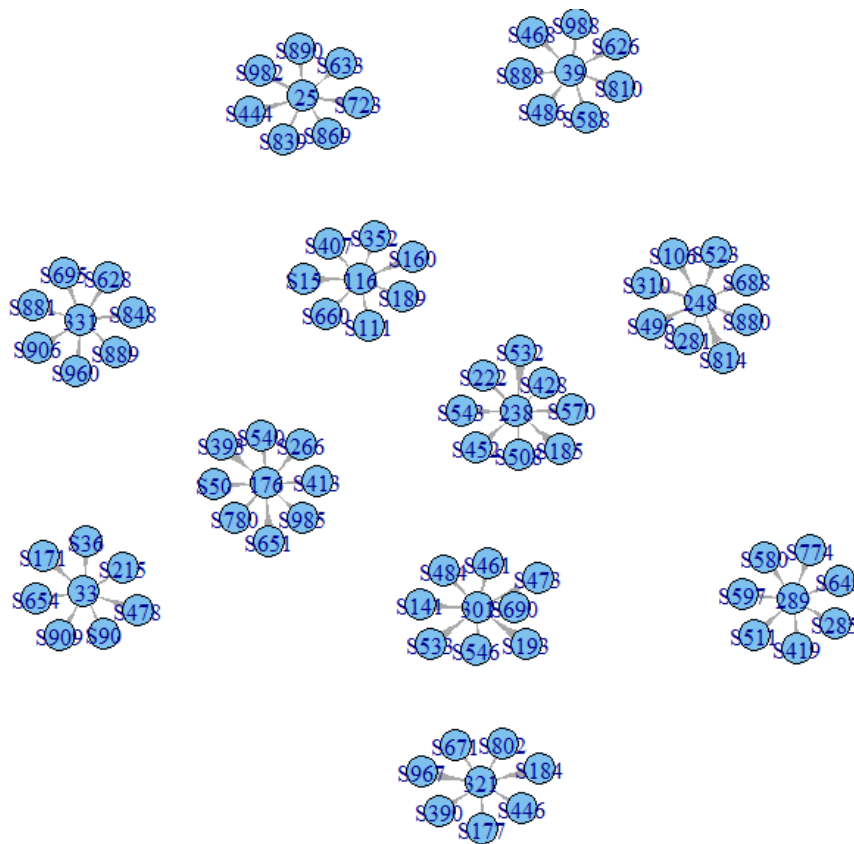


Figure 11. Graphical presentation of households containing more than 6 persons

2.3.3. Dubious objects

Checking for dubious objects is the same as checking for inconsistent objects. Here the function “relation” can also be used. The only difference is the interpretation of the results. In our example, persons or households that are returned when checking for inconsistent objects do violate a rule that should always be satisfied. In terms of edit rules these rules are called “hard” edit rules. However, persons or households that are returned when checking for dubious objects are only suspicious. They are not necessarily wrong but deserve special attention. In terms of edit rules these rules are called “soft” edit rules.

2.3.4. Measurement error

The only “measurement error” method that can be implemented is the one counting the number of records that are marked as containing measurement errors (given that the data source holder marks these values). This method can simply be executed by calculating the sum over all marks. There is no need to include it in our package.

2.3.5. Inconsistent values

To check for non-logical relations between values of variables, we use the editrules-package (De Jonge, and van der Loo, 2011) recently developed at Statistics Netherlands. This package can, among others be used, to load editrules specified in a text file, to visualize these rules and to apply these rules to a data file. As an example, we created a text file with editrules as shown in Figure 12.

These rules can be visualized using the “plot” function of the aforementioned package. This leads to the result shown in Figure 13. Variable names are presented in blue circles, and editrules in yellow squares. The rules are named by the package, the names are presented in Figure 12 (in red).

```

#categorical edit rules
dat9: sex %in% c("male","female")
dat7: employed %in% c("yes","no")
dat8: province %in% c("Akkerwinde", "Grasmalen", "Nieuwekans", "Lommerdal", "Smeulde", "Stapelrade", "Vuilpanne")
dat6: education %in% c("no formal education", "primary", "lower secondary", "upper secondary", "bachelor", "master")

#numerical edit rules
num1: age<=120
num2: age>=0
num3: income<=2500
num4: income>=0

#mixed edit rules
mix5: if (age<18) income<=0
mix6: if (age>65) income<=0
mix7: if (age<16) education %in% c("no formal education", "primary", "lower secondary", "upper secondary")
mix8: if (employed=="no") income<=10

```

Figure 12. Example of a text file containing editrules

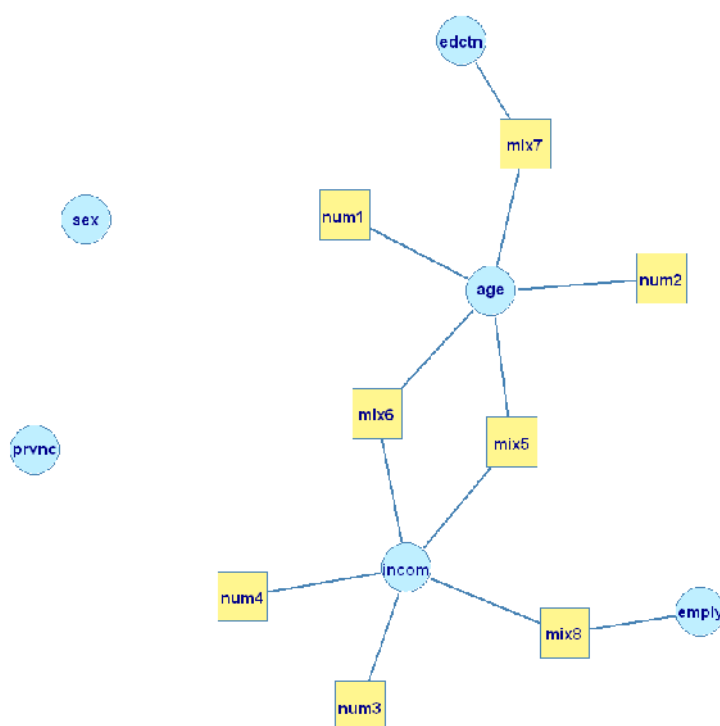


Figure 13. Visualization of editrules

Visualizing edit rules is particularly useful to gain insight into dependencies between rules, i.e. which variables are involved in more than one rule.

After applying the rules to the data file, the results can be visualized in a table as well as graphical using respectively the functions “summary”, and “plot” included in the package. An example of a table that shows the results is presented in Figure 14. The results show that editrule num3 (income<=2500) is violated most often, i.e. in 3,1% of the cases. The results furthermore show that for 95,9% of the records no editrules are violated. For 0,8% 4 editrules are violated.

```

> result<-violatedEdits(E, sam)
> summary(result)
Edit violations, 1005 observations, 0 completely missing (0%):

  editname freq  rel
    num3   31 3.1%
    dat7    8 0.8%
    mix7    2 0.2%

Edit violations per record:

  errors freq  rel
     0  964 95.9%
     1   33  3.3%
     4    8  0.8%
> |

```

Figure 14. Results (in table format) of applying editrules to a data file

2.3.6. Dubious values

Checking for dubious values is the same as checking for inconsistent values. The only difference is the interpretation of the results. When checking for inconsistent values “hard edit rules” are applied to the data, meaning that a violation of such a rule always implies that there is something wrong with the record. In case of dubious values, “soft edit rules” are applied to the data, meaning that a violation only indicates that the record is suspicious.

2.4. Completeness

Table 3 gives an overview of the measurement methods in the Completeness dimension, and the extent to which they can be implemented.

Table 3. Extent to which the indicators in the Completeness dimension can be implemented

Dimension indicators	Methods	Can be implemented?	Has been implemented?
3.1 Undercoverage	% of objects of the reference list missing in the source		 undercoverage
3.2 Overcoverage	% of objects in the source not included in the reference population		 overcoverage
	% of objects in the source not belonging to the target population of the NSI		 overcoverage
3.3 Selectivity	Use statistical data inspection methods, such as histograms, to compare a background variable (or more than one) for the objects in the data source and the reference population		 visualise
	Use of more advanced graphical methods, such as tableplots		 tableplot
	Calculate the R-indicator for the objects in the source		 R-script for R-indicators

			(Schouten, 2012)
3.4 Redundancy	% of duplicate objects in the source (with the same identification number)	✓	✓ redundancy
	% of objects in the source with the same values for a selection of variables	✓	✓ redundancy
	% of objects in the source with the same values for all variables	✓	✓ redundancy
3.5 Missing values	% of objects with a missing value for a particular variable	✓	✓ redundancy
	% of objects with all values missing for a selected (limited) number of variables	✓	✓ redundancy
	Use of graphical methods to inspect for missing values for variables	✓	✓ tableplot VIM package
3.6 Imputed values	% of imputed values per variable in the source	✓	✓
	Contact the data source holder and request the percentage of imputed values per variable		

The measurement methods are now discussed in more detail.

2.4.1. Undercoverage

The function “undercoverage” determines the coverage of a data source with respect to a reference dataset. An example of this function is provided in Figure 15. In this example the blood group register of Samplonia *reg_t1* is compared to the full population register of Samplonia. The person Id

```
> undercoverage(reg_t1, sam, "id_code", "id")
Source data contain 5 duplicated records. These records are omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id ( id_code )
Number of non-matching records: 600 (of 1000 )
Fraction of non-matching records: 0.6 , ( 60 %)
```

Figure 15. The function “undercoverage” can be used to determine the extent of undercoverage

(*Id_code* in *reg_t1*, and *Id* in *sam*) is used to link both files to each other. It appears that there is serious undercoverage, i.e. 600 of 1000 records in the population register do not match.

2.4.2. Overcoverage

Overcoverage can be explored using the function “overcoverage”. An example of this function is provided in Figure 16.

```
> overcoverage(sam, reg_t1, "id", "id_code")
Source data contain 5 duplicated records. These records are omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id ( id_code )
Number of non-matching records: 600 (of 1000 )
Fraction of non-matching records: 0.6 , ( 60 %)
```

Figure 16. The function “overcoverage” can be used to determine the extent of overcoverage

In this example the positions of the files *sam*, and *reg_t1* are switched compared to the undercoverage example. The result is exactly the same as the result shown in Figure 15. This is of course not surprising as overcoverage is the opposite of undercoverage .

By using this function both measurement methods:

- % of objects in the source not included in the reference population
- % of objects in the source not belonging to the target population of the NSI

can be executed. The only difference between both methods is the file that needs to be passed to the function, i.e. the reference population or the target population of the NSI.

2.4.3. Selectivity

Several graphical methods are implemented to check for selectivity. The most advanced one, i.e. the table plot, has already been shown in Figure 5. Another option is to use the function “visualise” with two files as input (in Figure 4 an example with one file as input has already been discussed). Two examples of results having the files *sam*, and *reg_t1* as input are shown in Figure 17, and Figure 18. In Figure 17 all variables are shown that are at least in one of the two files. For the three variables that are in both files grouped histograms are drawn. In this way the distributions of the values can be compared between the two files. For the variables only occurring in one file standard histograms are shown.

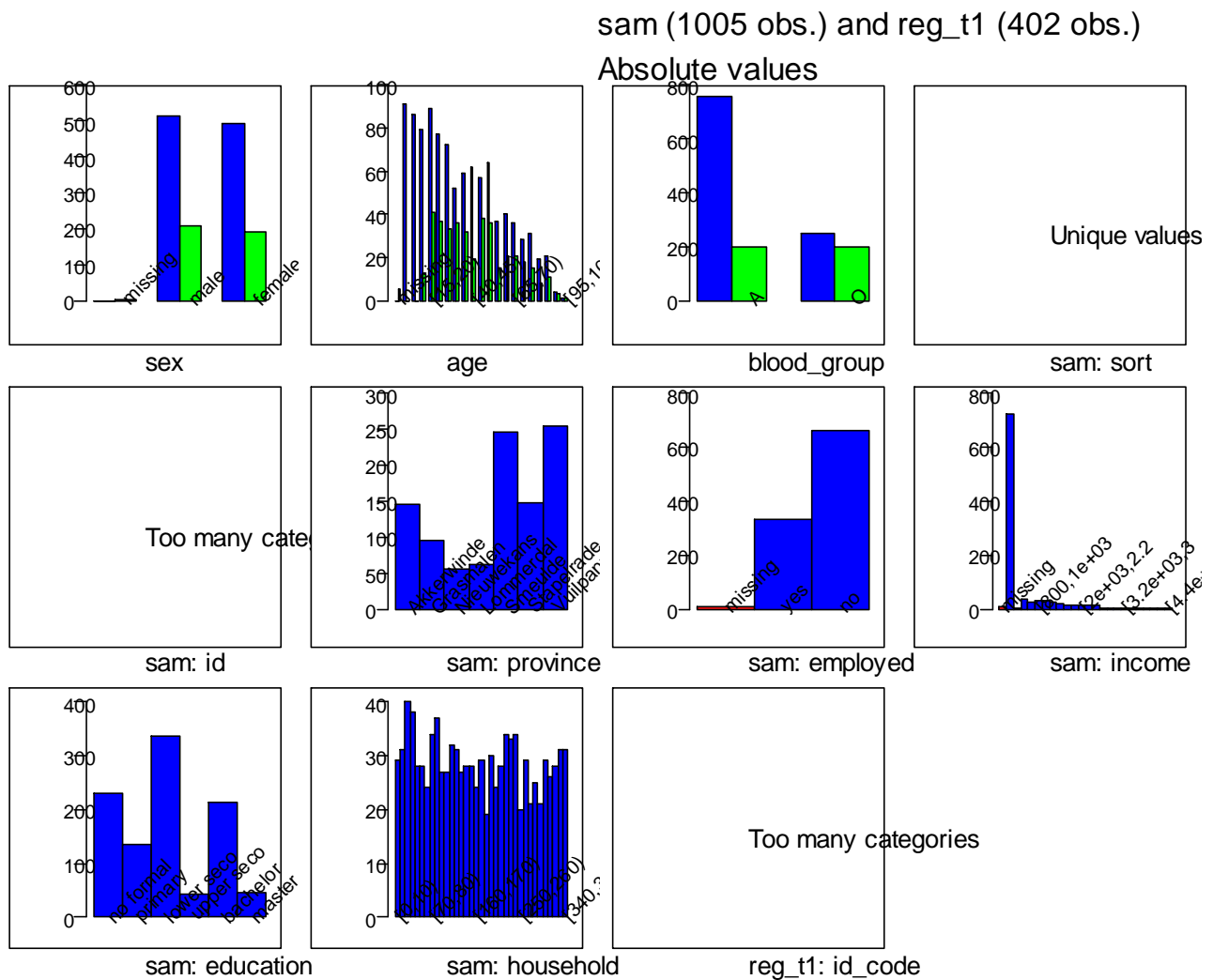


Figure 17. The function “visualise” can be used to explore for selectivity. In this example all values are plotted in an absolute way

Comparing the distributions can be difficult when there is a large difference between the number of records in both files (see Figure 17 for an example). The bars are in this case much higher for one file than for the other file. It can therefore be easier to show the relative shares of all categories in the total number of records in the file, i.e.

$$\frac{\text{number of records in 1 category}}{\text{total number of records}}$$

These relative shares are shown in Figure 18. From this example it appears that the lower age classes are clearly underrepresented in the blood group register. Also the distributions over the blood groups seem to differ between both files. The distribution over the sexes looks comparable between the files.

Alternatively, if sufficient background information is available the R-indicator can be calculated (De Heij *et al.*, 2010) using the scripts created for that purpose (Schouten, 2012).

2.4.4. Redundancy and missing values

The measurement methods used for checking for redundancy and missing values can all be executed using the function “redundancy”. Examples of the output of this function are shown in Figure 19. The output consists of two parts. In the upper part a table is shown in which the missing, unique, and duplicated values per variable are displayed. In the lower part all variables are considered together, i.e. how many records do have the same values for all variables, for how many

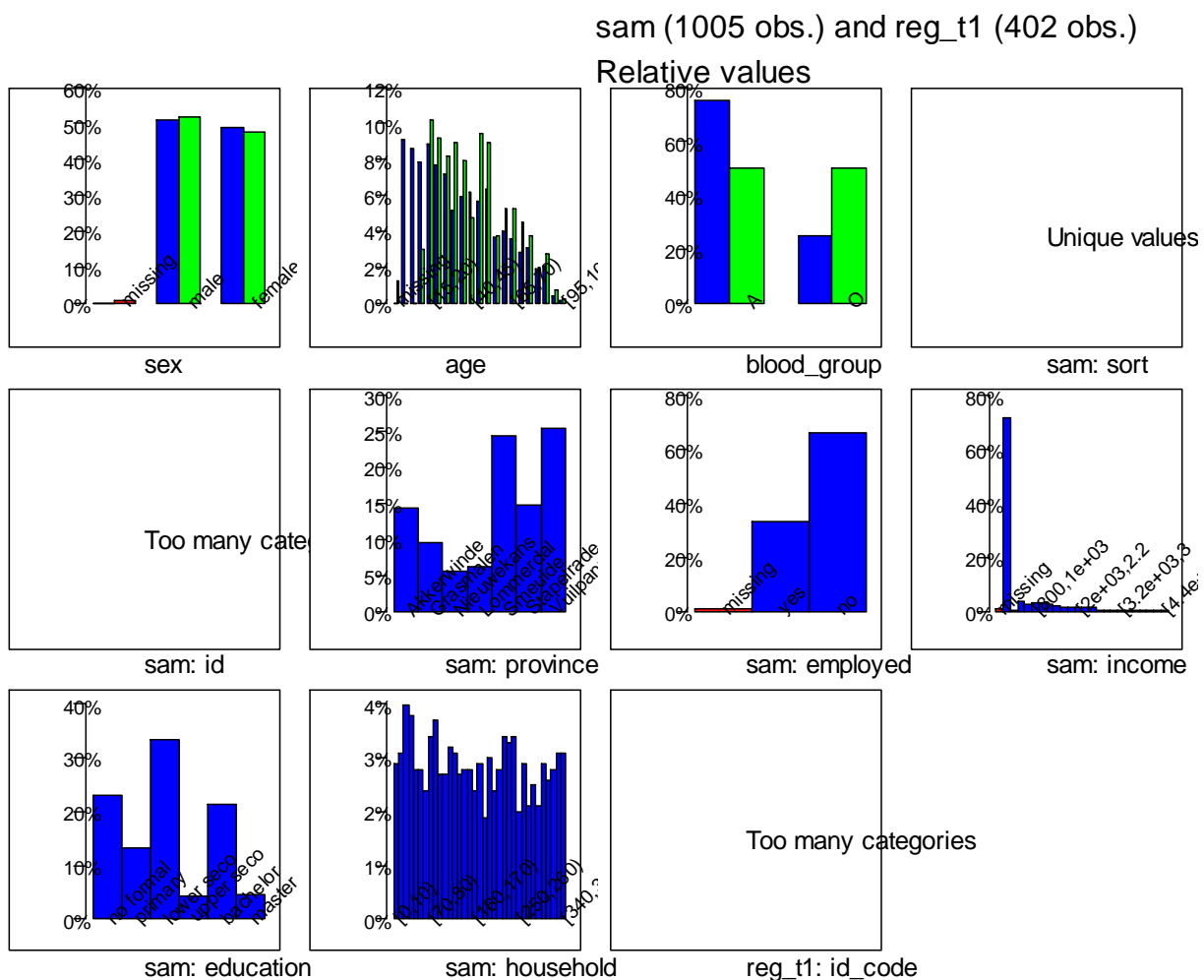


Figure 18 The function “visualise” can be used to explore for selectivity. In this example all values are plotted in a relative way

records are the values for the considered variables completely or partly missing, and so on. By adding the argument “percentage=TRUE” to the function also percentages are calculated. The variables to be evaluated are specified by the user. In the first example in Figure 19 no variables are specified such that the function implicitly considers all variables. In the second example the variables Sex, Age, Employed, and Income are selected for being evaluated.

```
> redundancy(sam)
summary of values per variable
      values missing unique duplicated
sort      1005      0    1005         0
id         1005      0    1000         5
province   1005      0      7         998
sex         1005      0      2        1003
age         1005      0     94         911
employed   997      8      2         995
income     997      8    310         687
education  1005      0      6         999
blood_group 1005      0      2        1003
household  1005      0    324         681
-----
summary of objects regarding all selected variables
      total      complete partly_missing completely_missing unique
duplicated
      1005      997          8              0          1005
      0
> redundancy(sam, vars = c("sex", "age", "employed", "income"))
summary of values per variable
      values missing unique duplicated
sex      1005      0      2        1003
age      1005      0     94         911
employed 997      8      2         995
income   997      8    310         687
-----
summary of objects regarding all selected variables
      total      complete partly_missing completely_missing unique
duplicated
      1005      997          8              0          495
      510
```

Figure 19. Example of the output of the function “redundancy”

All specified methods for redundancy can be performed using this function as they only differ with regard to the columns that should be passed to the function. In illustration:

- For the measurement method “% of duplicate objects in the source (with the same identification number)” *all columns that are part of the primary key* should be considered. The result can then be read in the lower part of the output.
- For the measurement method “% of objects in the source with the same values for a selection of variables” *a selection of columns* should be passed to the function and evaluated. The result can then be read in the lower part of the output.
- For the measurement method “% of objects in the source with the same values for all variables” *all columns* should be evaluated. The result can then be read in the lower part of the output.

The same holds for the methods that are used for checking for missing values:

- For the measurement method “% of objects with a missing value for a particular variable” at least the variables of interest should be passed to the function, and the result can be read from the upper part of the output.
- For the measurement method “% of objects with all values missing for a selected (limited) number of variables” *a selection of columns* should be passed to the function and evaluated. The result can then be read in the lower part of the output.

The R package VIM (Templ *et al.*, 2012) offers various graphical methods to analyze missing values. One of them is also implemented in the function “compare” (see Figure 8). Missing values can also be studied graphically with a table plot as shown in Figure 5.

2.4.5. Imputed values





The only “measurement error” method that can be implemented for imputed values is the one counting the number of records that are marked as imputed (given that the data source holder marks these values). This method can simply be executed by calculating the sum over all marks and is therefore not included in our package.

2.5. Time-related

Table 4 gives an overview of the measurement methods in the Time-related dimension, and the extent to which they can be implemented.

Table 4. Extent to which the indicators in the time-related dimension can be implemented

Dimension indicators	Methods	Can be implemented?	Has been implemented?
4.1 Timeliness	Time difference (days) = (Date of receipt by NSI) – (Date of the end of the reference period over which the data source reports)	✓	✓ dateDiff
	Time difference (days) = (Date of receipt by user) – (Date of the end of the reference period over which the data source reports)	✓	✓ dateDiff
4.2 Punctuality	Time difference (days) = (Date of receipt by NSI) – (Date agreed upon; as laid down in the contract)	✓	✓ dateDiff
4.3 Overall time lag	Total time difference (days) = (Predicted date at which the NSI declares that the source can be used) – (Date of the end of the reference period over which the data source reports)	✓	✓ dateDiff
4.4 Delay	Contact the data source holder to provide their information on registration delays		
	Time difference (days) = (Date of capturing the change in the source by the data source holder) – (Date the change occurred in the population)	✓	✓ dateDiff
4.5 Dynamics of objects	% Births t = (Births t / Total objects t) x 100% = (Births t / (Births t + Alive t)) x 100%	✓	✓ birth
	% Deaths t-1 = (Deaths t / Total objects t-1) x 100% = (Deaths t / (Alive t + Deaths t)) x 100%	✓	✓ death1 death2

	$\text{Change in objects} = \frac{((\text{Births } t + \text{Alive } t) - (\text{Alive } t + \text{Deaths } t))}{(\text{Alive } t + \text{Deaths } t) \times 100\%}$ $= \frac{(\text{Births } t - \text{Deaths } t)}{(\text{Alive } t + \text{Deaths } t) \times 100\%}$	✓	 change change_fast
4.6 Stability of variables	Use statistical data inspection methods to compare the values of specific variables for persistent objects in different deliveries of the source. Graphical methods that can be used are a bar plot and a scatter plot	✓	 compare visualise
	$\% \text{ of Changes} = \frac{(\text{Number of objects with a changed value} / \text{total number of persistent objects with a value filled in for the variable under study}) \times 100\%}{}$	✓	 changed_value unchanged_value
	A correlation statistical method can be used to determine to which extent values changed in the same direction for different objects. For categorical data a method such as Cramér's V can be used	✓	 cor (standard R function) cramerV.test

The measurement methods are now discussed in more detail.

2.5.1. Timeliness, Punctuality, Overall time lag, Delay

The measurement methods:

- Time difference (days) = (Date of receipt by NSI) – (Date of the end of the reference period over which the data source reports)
- Time difference (days) = (Date of receipt by user) – (Date of the end of the reference period over which the data source reports)
- Time difference (days) = (Date of receipt by NSI) – (Date agreed upon; as laid down in the contract)
- Total time difference (days) = (Predicted date at which the NSI declares that the source can be used) – (Date of the end of the reference period over which the data source reports)
- Time difference (days) = (Date of capturing the change in the source by the data source holder) – (Date the change occurred in the population)

all have in common that the difference between two dates has to be determined. This can be done using the function “dateDiff”. This function returns the difference between two dates in various formats as can be seen in Figure 20. When only one date is entered, the function determines the difference with the current date, i.e. the date at which the measurement method is executed.

```
> dateDiff("15-01-2012", "16-01-2012")
Time difference of 1 days
> dateDiff("15-01-2012", "17-02-2012", unit = "weeks")
Time difference of 4.714286 weeks
> dateDiff("15-01-1998", "16-01-2012", unit = "months")
Time difference of 168 months
> dateDiff("15-01-1998", "16-01-2012", unit = "years")
Time difference of 14 years
```

Figure 20. Illustration of the working of the function “dateDiff”

2.5.2. Dynamics of objects

To explore population dynamics several functions have been implemented, i.e. “birth”, “death2”, “death1”, “change”, and “change_fast”. Examples of these functions are shown in Figure 21.

```
> birth(reg_t1, reg_t2, "id_code")  ## in t2 but not in t1, overcoverage t2 with number of records at
t2 as base
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Number of new records: 3 (of 402 )
Fraction of new records: 0.007462687 , ( 0.7462687 %)
> death2(reg_t1, reg_t2, "id_code")  ## undercoverage at t2 (records in t1 that are absent in t2)
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Number of removed records: 2 (of 402 )
Fraction of removed records: 0.004975124 , ( 0.4975124 %)
> death1(reg_t1, reg_t2, "id_code")  ## undercoverage at t2 with number of records at t1 as base
(records in t1 that are absent in t2)
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Number of removed records: 2 (of 401 )
Fraction of removed records: 0.004987531 , ( 0.4987531 %)
> change(reg_t1, reg_t2, "id_code")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Number of new records: 3 , number of removed records: 2
Overall change in records: 1 (of 402 )
Fraction of change in records: 0.002487562 , ( 0.2487562 %)
> change_fast(reg_t1, reg_t2, "id_code")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Overall change in records: 1 (of 402 )
Fraction of change in records: 0.002487562 , ( 0.2487562 %)
>
```

Figure 21. Examples of functions that are used to explore population dynamics

The function “birth” determines the fraction of records that are only contained in the second time period (*reg_t2* in the example):

$$\frac{\text{Births } t_2}{\text{Total number of objects at } t_2}$$

According to Figure 21, 3 new records appear in *reg_t2* that were not yet present in *reg_t1*.

The functions “death2”, and “death1” are used to determine the fraction of records present in the first time period (*reg_t1*) but absent in the second time period (*reg_t2*). The difference between the functions is that “death2” takes the second time period as reference time period, while “death1” takes the first period as reference:

$$\text{Death2} = \frac{\text{Deaths } t_2}{\text{Total number of objects at } t_2}, \text{Death1} = \frac{\text{Deaths } t_2}{\text{Total number of objects at } t_1}$$

This also explains the difference between the outcomes of “death2”, and “death1” in the example of Figure 21. The file *reg_t1* contains 401 observations (402 minus 1 duplicated record), while the file *reg_t2* contains 402 observations (403 minus 1 duplicated record).

The functions “change”, and “change_fast” provide the fraction of records that changed between t_1 and t_2 . The difference between both functions is that “change” also returns a summary of new and

removed records as can be seen in Figure 21. This takes up more time, which is the reason ‘fast’ is included in the less calculation intensive method.

2.5.3. Stability of variables

The function “compare” can be used to visually compare the values of individual records between different time periods (in Figure 8 this function was already used to compare the values for the same variable in two different files referring to the same time period). To compare the distributions of values between two different time periods the function “visualise” can be applied (Figure 17, and Figure 18 show examples of comparing two files referring to the same period).

Two functions have been implemented to check for the number of (un)changed values: “changed_value”, and “unchanged_value”. Results for these functions when comparing the files *reg_t1*, and *reg_t2* are shown in Figure 22.

```
> changed_value(reg_t1, reg_t2, keys="id_code", variable = "age")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Variable compared: age
Number of records changed: 0 (of 399 aligned records)
Fraction of records changed: 0 , ( 0 %)
> changed_value(reg_t1, reg_t2, keys="id_code", variable = "sex")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Variable compared: sex
Number of records changed: 0 (of 399 aligned records)
Fraction of records changed: 0 , ( 0 %)
> changed_value(reg_t1, reg_t2, keys="id_code", variable = "blood_group")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Variable compared: blood_group
Number of records changed: 0 (of 399 aligned records)
Fraction of records changed: 0 , ( 0 %)
> unchanged_value(reg_t1, reg_t2, keys="id_code", variable = "age")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Variable compared: age
Number of records unchanged: 399 (of 399 aligned records)
Fraction of records unchanged: 1 , ( 100 %)
> unchanged_value(reg_t1, reg_t2, keys="id_code", variable = "sex")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Variable compared: sex
Number of records unchanged: 399 (of 399 aligned records)
Fraction of records unchanged: 1 , ( 100 %)
> unchanged_value(reg_t1, reg_t2, keys="id_code", variable = "blood_group")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 1 duplicated record. This record is omitted.
Keys used: id_code
Variable compared: blood_group
Number of records unchanged: 399 (of 399 aligned records)
Fraction of records unchanged: 1 , ( 100 %)
>
```

Figure 22. Examples of functions “changed_value”, and “unchanged_value”

The function “changed_value” determines the number of values in a data source that changed between two time instances. The function “unchanged_value” determines the number of unchanged values. According to Figure 22 no values changed for the variables Age, Sex, and Blood_group.















To calculate the correlation in order to determine to which extent values changed in the same direction the standard R function “cor” can be used. Cramér’s V can be calculated using the function “cramerV.test” which has been incorporated in the dataquality package.

2.6. Integrability

Table 5 gives an overview of the measurement methods in the Integrability dimension, and the extent to which they can be implemented.

Table 5. Extent to which the indicators in the integrability dimension can be implemented

Dimension indicators	Description	Can be implemented?	Has been implemented?
5.1 Comparability of objects	% of identical objects = (Number of objects with exactly the same unit of analysis and same concept definition as those used by NSI / Total number of relevant objects in source) x 100%		 coverage
	% of corresponding objects = (Number of objects that, after harmonization, would correspond to the unit needed by NSI / Total number of relevant objects in source) x 100%		
	% of incomparable objects = (Number of objects that, even after harmonization, will not be comparable to one of the units needed by NSI / Total number of relevant objects in source) x 100%		
	% of non-corresponding aggregated objects = (% of objects of interest at an aggregated level in source 1 + % of objects of interest at the same aggregated level in source 2)		
5.2 Alignment of objects	% of identical aligned objects = (Number of objects in the business register with exactly the same unit of analysis and same concept definition as those in the source / Total number of relevant objects in business register) x 100%		 coverage
	% of corresponding aligned objects = (Number of objects in the business registers that, after harmonization, correspond to units or parts of units in the source / Total number of relevant objects in business register) x 100%		

	<p>% of non-aligned objects = (Number of objects in the business register that, even after harmonization of the objects in the source, can not be aligned to one of the units in the source / Total number of relevant objects in business register) x 100%</p>		
	<p>% of non-aligned aggregated objects = (% of objects of interest at an aggregated level in source 1 that can not be aligned + % of objects of interest at the same aggregated level in source 2 that can not be aligned)</p>		
5.3 Linking variable	<p>% of objects with no linking variable = (Number of objects in source without a linking variable / Total number of objects in the source) x 100%</p>		 redundancy
	<p>% of objects with (a) linking variable(s) different from the one(s) used by NSI = (Number of object in source with (a) linking(s) variable different from the one used by the NSI / Total number of objects with (a) linking variable(s) in the source) x 100%</p>		 checkFormat
	<p>% of objects with correctly convertible linking variable(s) = (Number of objects in the source for which the original linking variable can be converted to one used by the NSI / Total number of objects with a linking variable in the source) x 100%</p>		
5.4 Comparability of variables	<p>Use statistical data inspection methods to compare the values of objects or totals of object aggregates for variables in both sources. Graphical methods that can be used are a bar plot and a scatter plot. Distributions of values can also be compared.</p>		 compare visualise
	<p>The Mean Absolute Percentage Error (MAPE). MAPE has a lower bound of zero but has no upper bound. Alternatively the symmetric MAPE could be used. This method measures the symmetric mean of the absolute percentage error were the</p>		 MAPE sMAPE

deviation between the percentage distributions is divided by the half-sum of the deviations.		
A method derived from the chi-square test that evaluates the distributions of the numeric values in both data sets. For categorical data Cramér's V (Cramér, 1946) could be used.	✓	chisq.test (standard R function) cramerV.test
% of objects with identical variable values = (Number of objects in source 1 and 2 with exactly the same value for the variable under study / Total number of relevant objects in both sources) x 100%	✓	changed_value unchanged_value

The measurement methods are now discussed in more detail.

2.6.1. Comparability of objects

The measurement method “% of identical objects” can be determined using the function “coverage”. An example of this function is shown in Figure 23. In this example 400 of the 401 (unduplicated) records can be linked to the reference file *sam*.

```
> coverage(reg_t1, sam, "id_code", "id")
Source data contain 1 duplicated record. This record is omitted.
Reference data contain 5 duplicated records. These records are omitted.
Keys used: id_code ( id )
Number of matching records: 400 (of 401 )
Fraction of matching records: 0.9975062 , ( 99.75062 %)
>
```

Figure 23. Illustration of the function “coverage”

The measurement method “% of corresponding objects = (Number of objects that, after harmonization, would correspond to the unit needed by NSI / Total number of relevant objects in source) x 100%” can partly be evaluated using the function “coverage”. The harmonization method itself is not implemented in the package as it does very strongly depend on the statistic at hand. However, after harmonization, the function “coverage” can be applied to the harmonized data.

The “% of incomparable objects” is comparable to the “% of corresponding objects”, the only difference is that now the fraction of records is considered that even after harmonization can not be linked to the records of the NSI. This fraction can therefore be calculated by performing the following calculation:

$$\% \text{ of incomparable objects} = 100\% - \% \text{ of corresponding objects}$$

To determine the “% of non-corresponding aggregated objects” first an aggregation has to be performed to the data after which the function “coverage” can be applied. The aggregation itself is implemented in several base R functions (tapply, and aggregate) as well as in several R packages (such as plyr and data.table). Experienced R users can easily aggregate their data in R and pass them to the function “coverage”.

2.6.2. *Alignment of objects*

The measurement methods for this indicator are comparable to the measurement methods for the previous indicator (section 2.6.1). The only difference is that the reference file is now the Business Register.

2.6.3. *Linking variable*

The “% of objects with no linking variable” can be determined using the earlier discussed function “redundancy” (section 2.4.4). This function can easily be used to determine the percentage of objects for which values for the variables in the primary key are (partly) missing.

The “% of objects with (a) linking variable(s) different from the one(s) used by NSI” can be used by applying the earlier discussed function “checkFormat” (section 2.3.1). This function checks whether a linking variable is in line with the format needed by the NSI.

The “% of objects with correctly convertible linking variable(s)” can partly be determined by the package. The conversion itself is not part of the package as it is strongly statistic specific. After conversion the function “checkFormat” can be applied.

2.6.4. *Comparability of variables*

The measurement method “Use statistical data inspection methods to compare the values of objects or totals of object aggregates for variables in both sources” has in fact already been discussed in section 2.5.3. In section 2.5.3 the “Stability of variables” has been discussed. This implies that variable values for the same data source but at different time instants are compared. The only difference with “Comparability of variables” is that for the last indicator two different data sources are compared instead of one data source at two time instants. The same functions (“compare”, and “visualise”) can be used for both indicators. The only difference is that different files are passed to the functions, i.e. to test for stability “file data source A at t_1 ”, and “file data source A at t_2 ” and to test for comparability “file data source A at t_1 ”, and “file data source B at t_1 ”.

To the measurement method “% of objects with identical variable values” the same reasoning can be applied. Like for the stability analysis the functions “changed_value”, and “unchanged_value” can be used. The only difference is the files that are passed to the function.

The Mean Absolute Percentage Error (MAPE) can be determined using the functions “MAPE”, and “sMAPE”. Cramér’s V can be calculated using the function “cramerV.test”. Both functions have been incorporated in the dataquality package. A chi-square test can be done using the standard R function “chisq.test”.

3. **The Quality Report Card**

To report the evaluation findings of the quality components identified for administrative data a quality report card has been developed. Its full name is Quality Report Card for Administrative data (QRCA) and it is used to standardize the reporting of the evaluation results obtained. The report card is included at the end of this document. Both the findings of the automated (R-scripted) methods described above and the non-automated methods can be noted in the QRCA.

The user starts filling in the QRCA by reporting the results for the Technical Checks dimension, followed by the Time-related, Completeness, Accuracy and Integrability dimension findings. For each dimension, the results of all the measurements methods performed can be noted down. Next, these findings must be converted to a score at the indicator level. For these more general scores the signs +, o, - and ? are proposed for good, reasonable, poor and unclear. Intermediary scores are created by combining symbols with a slash (/) as a separator. This is identical to the scores used in

the metadata-checklist of Daas and Ossen (2011). Additional space is included to write down remarks. For each dimension, the evaluation ends with an overall conclusion. This needs to be filled-in by the user. Possible dimensional findings are 'No problems found' (green checkbox), 'Some minor issues observed' (orange checkbox) and 'Serious problems detected' (red checkbox). When serious problems are found, this should be noted and evaluation must stop. Any instructions included should be followed. When minor issues are found, this should also be noted but evaluation may continue. When no problems are found evaluation should also continue. When serious issues are found or when all dimensions have been evaluated, the general findings section should be filled in. First, the scores found at the dimensional level are copied. Next a user has the option to provide additional information at the object or variable level for each dimension; if applicable. Additional space is included to write down remarks. The summary ends with an overall conclusion for the quality of the data studied which has to be filled in by the user.

References

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

Daas, P.J.H., Ossen, S.J.L. (2011), Metadata Quality Evaluation of Secondary Data Sources. *International Journal for Quality Research*, 5 (2), 57-66.

Daas, P., Ossen, S., Tennekes, M., Zhang, L-C., Hendriks, C., Foldal Haugen, K., Bernardi, A., Cerroni, F., Laitila, T., Wallgren, A., Wallgren, B. (2011a), List of quality groups and indicators identified for administrative data sources, First deliverable of WP4 of the BLUE-ETS project, March 10.

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. (2011b) Report on methods preferred for the quality indicators of administrative data sources. Second deliverable of workpackage 4 of the BLUE Enterprise and Trade Statistics project, September 28.

De Jonge, E., van der Loo, M. (2011). Editrules: R package for parsing and manipulating edit rules. R package version 2.5-0. <http://code.google.com/p/editrules>

De Heij, V., Schouten, B., Shlomo, N. (2010). RISQ manual: Tools in SAS and R for the computation of R-indicators and partial R-indicators, Work package 8, Deliverable 12.1. <http://www.risq-project.eu/papers/RISQ-Deliverable-12-1.pdf>

James, D. A. (2011). RSQLite: SQLite interface for R. R package version 0.11.1. <http://CRAN.R-project.org/package=RSQLite>

Ripley, B., Lapsley, M. (2012). RODBC: ODBC Database Access. R package version 1.3-6. <http://CRAN.R-project.org/package=RODBC>

Schouten, B. (2012) Webpage: "RISQ: Representative Indicators for Survey Quality." <http://www.risq-project.eu/tools.html>. Accessed June 2012.

Templ, M., Alfons, A., Kowarik A., Prantner B. (2012). VIM: Visualization and Imputation of Missing Values. R package version 3.0.0. <http://CRAN.R-project.org/package=VIM>

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

Tennekes, M. and De Jonge, E. (2012). Tabplot: Tableplot, a visualization of large datasets. R package version 0.11-2. <http://CRAN.R-project.org/package=tabplot>

Tennekes, M., de Jonge, E., Daas, P.J.H. (2013) Visualizing and Inspecting Large Datasets with Tableplots. *Journal of Data Science*, 11 (1), 43-58.

Van der Laan, J. (2012). LaF: Fast access to large ASCII files. R package version 0.4. <http://CRAN.R-project.org/package=LaF>

Quality Report Card for Administrative data



EUROPEAN COMMISSION
European Research Area



Funded under Socio-economic Sciences & Humanities

Quality Report Card for Administrative data (QCRA)



Version 0.4

Data source studied

--	--

Document Version

Version	Adaptations	Responsible	Date
0.2	Based on the initial idea and the proposal included in BLUE-ETS WP4 deliverable 2 a Quality Report Card for Administrative data is proposed	Statistics Netherlands	22/06/2012
0.3	Adjusted version as result of WP8 case studies	Statistics Netherlands	18/01/2013
0.4	Adjustments as result of manual creation	Statistics Netherlands	25/01/2013

Document description

This document includes a Quality report Card for Administrative data used for the evaluation of the statistical usefulness of administrative data. The list focuses on the essential quality dimensions and indicators identified in deliverables 1 and 2 of BLUE-ETS workpackage 4. Evaluation is performed by a user at a National Statistical Institute (NSI). The outcome of the evaluation assists the user in the decision to use the data for statistics production.

Instructions

To report the evaluation findings of the quality indicators identified for administrative data a quality report card has been developed. It is named a Quality Report Card for Administrative data (QRCA) and is used to standardize reporting of the evaluation results obtained. Both the findings of automated (scripted) and non-automated quality indicators methods can be noted in the QRCA. The user starts filling in the QRCA by reporting the results for the Technical Checks dimension, followed by the Time-related, Completeness, Accuracy and Integrability dimension findings. For each dimension, the results of all the measurement methods that can be performed have to be noted down. Quantitative scores can be noted directly. Qualitative scores need to be expressed by the signs +, o, - and ? which are used to identify good, reasonable, poor and unclear. Intermediary scores are created by combining symbols with a slash (/) as separator. This is identical to the scores used in the metadata-checklist of Daas and Ossen (2011). Additional space is included to write down remarks.

For each dimension, the evaluation ends with an overall conclusion. This needs to be filled-in by the user. Possible dimensional findings are ‘No problems found’ (green checkbox), ‘Some minor issues observed’ (orange checkbox) and ‘Serious problems detected’ (red checkbox). When serious problems are found, this should be noted and –very likely- evaluation must stop. Any instructions included should be followed. When minor issues are found, this also needs to be noted but evaluation can continue. When no problems are found evaluation should also continue of course. In case of serious issues or when all dimensions have been evaluated, the general findings section should be filled in. First, the scores found at the dimensional level are copied and (if needed) converted to the signs +, o, - and ? (see above). If needed, the user can provide additional information at the object or variable level for each dimension. Here also, additional space is included to write down remarks. The summary ends with an overall conclusion for the quality of the data studied which has to be filled in by the user. Plots or other graphical representations of findings can be added to the document if needed.

Information of the NSI-representative who fills in the report

	Full name	
	Position	
	Department	
	Phone number	
	E-mail address	
	Date on which the report card was completed	

1. Technical checks

The symbols used to indicate the scores are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combining symbols with a slash (/) as a separator.

	Indicators	Level	Score	Remarks
1.1	Readability			
1.2	File declaration compliance			
1.3	Convertability			

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
	No problems found	☐
	Some minor issues observed	☐
	Serious problems detected	☐
Write additional remarks here:		

- Only continue when the GREEN or ORANGE marked area is checked.
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted and the section or person responsible for receipt of the data needs to be contacted.

2. Integrability

The symbols used to indicate the scores are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combining symbols with a slash (/) as a separator.

	Indicators	Level	Measurement method results	Score	Remarks
2.1	Comparability of objects	Objects			
2.2	Alignment of objects	Objects			
2.3	Linking variable	Variables			
2.4	Comparability of variables	Variables			

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
	No problems found	<input type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here:		

- Only continue when the GREEN or ORANGE marked area is checked.
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.

3. Accuracy

The symbols used to indicate the scores are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combining symbols with a slash (/) as a separator.

	Indicators	Level	Measurement method results	Score	Remarks
3.1	Authenticity	Objects			
3.2	Inconsistent objects	Objects			
3.3	Dubious objects	Objects			
3.4	Measurement error	Variables			
3.5	Inconsistent values	Variables			
3.6	Dubious values	Variables			

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
	No problems found	<input type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here:		

- Only continue when the GREEN or ORANGE marked area is checked.
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.

4. Completeness

The symbols used to indicate the scores are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combining symbols with a slash (/) as a separator.

	Indicators	Level	Measurement method results	Score	Remarks
4.1	Undercoverage	Objects			
4.2	Overcoverage	Objects			
4.3	Selectivity	Objects			
4.4	Redundancy	Objects			
4.5	Missing values	Variables			
4.6	Imputed values	Variables			

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
	No problems found	<input type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here:		

- Only continue when the GREEN or ORANGE marked area is checked.
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.

5. Time-related dimension

The symbols used to indicate the scores are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combining symbols with a slash (/) as a separator.

	Indicators	Level	Measurement method results	Score	Remarks
5.1	Timeliness				
5.2	Punctuality				
5.3	Overall time lag				
5.4	Delay				
5.5	Dynamics of objects	Objects			
5.6	Stability of variables	Variables			

Dimensional findings

Briefly describe the overall findings for this dimension and (if required) the action that needs to be taken.

Overall conclusion	Dimensional score	
	No problems found	<input type="checkbox"/>
	Some minor issues observed	<input type="checkbox"/>
	Serious problems detected	<input type="checkbox"/>
Write additional remarks here:		

- Only continue when the GREEN or ORANGE marked area is checked.
When an ORANGE marked area is checked, make sure that these findings are noted as remarks or additional remarks.
- When the RED marked area is checked evaluation needs to be halted. If this problem can't be solved it needs to be concluded that the data is not suited for use by the NSI.

General findings

Data source studied

--	--

Dimensional scores

The symbols used to indicate the scores are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combing symbols with a slash (/) as a separator.

	Data dimensions	Level	Score	Remarks
1	Technical Checks	Overall		
2	Integrability	Overall		
		Objects		
		Variables		
3	Accuracy	Overall		
		Objects		
		Variables		
4	Completeness	Overall		
		Objects		
		Variables		
5	Time related	Overall		
		Objects		
		Variables		

Overall conclusion	Overall score	
<i>Write additional remarks here:</i>	<i>Negative</i>	<input type="checkbox"/>
	<i>Neutral</i>	<input type="checkbox"/>
	<i>Positive</i>	<input type="checkbox"/>