Evaluating administrative data quality as input of the statistical production process[1]

Fulvia Cerroni[2]*, Grazia Di Bella*[3]*, Lorena Galiè* [4]

Sommario

Valutare e analizzare la qualità dei dati da fonte amministrativa è un'esigenza crescente negli Istituti nazionali di statistica (INS), poichè i processi di produzione sempre più utilizzano questo tipo di dati. Monitorare la qualità delle forniture di dati amministrativi che entrano nel processo di produzione statistica, valutare il loro possibile utilizzo a fini statistici, supportare l'acquisizione dei dati amministrativi sono azioni che devono essere eseguite nelle prime fasi del processo di produzione. L'articolo riporta l'esperienza dell'Istat nell'ambito del progetto europeo BLUE-ETS volto a sviluppare un quadro concettuale della qualità dei dati amministrativi sulla base di un approccio multidimensionale di indicatori di qualità e a definire un nuovo strumento, la Quality Report Card, che possa essere associato ai dati amministrativi e utilizzato in generale dagli INS.

Parole chiave: dati amministrativi, qualità dei dati, indicatori di qualità dell'input.

Abstract

Evaluating and reporting data quality of administrative sources is a growing need for National Statistical Institutes (NSIs) as more and more production processes are using this type of data source. To monitor the quality of the administrative data supply that enter the statistical production process, to evaluate its possible use for statistical purposes and to support administrative data acquisition are tasks that should be performed in the early stages of the production process. The paper reports Istat experience within the European project BLUE-ETS in developing a conceptual framework of administrative data quality based on a multidimensional approach of quality indicators and in defining a new comprehensive instrument, the Quality Report Card, that can be associated to administrative data and generally used by NSIs.

Keywords: administrative data, data quality, input quality indicators.

[1] The authors thank Piet Daas and the other BLUE ETS WP4 members. Although the paper is the result of the combined work of the authors, the Sections are to be awarded as it follows: Sections 1, 2.1, 4 and 5 to Grazia Di Bella; Sections 2.1 and 2.2 to Cerroni Fulvia; Section 3 to Lorena Galiè. The text published exclusively engages the authors, the views expressed do not imply any liability by Istat.

[2] Technical assistant (Istat), e-mail: cerroni@istat.it.

[3] Senior researcher (Istat), e-mail: dibella@istat.it.

[4] Technical assistant (Istat), e-mail: galie@istat.it.

The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

1. The need to define the statistical quality of administrative data

Administrative data (AD) were added in the last few years as a further source next to the data collected from sample and census surveys, for the production of official statistics in National Statistical Institutes (NSIs).

This synergy enables to make the statistical production process more efficient: it is possible to expand the available statistical information, to reduce the so-called "statistical burden" among economic agents, to maximize available resources (Unece, 2007). But, if the use of administrative sources in the statistical process allows to reduce the task of data capturing, new tools need to be addressed: a) for the acquisition of AD; b) for their use in the statistics production process. From the management point of view, AD acquisition requires to define new production functions and new organizational structures capable of maintaining relations with AD holders (ADH) and improving the collaboration process. From the methodological point of view it is necessary to define shared and standardized procedures to meet *a posteriori* the quality standards imposed by official statistics (Wallgren and Wallgren, 2007).

International organizations, involved in producing statistics, are favoring the development of standardized methodologies or the sharing of best practices for the use of AD and there are many initiatives on this subject[5]. Considering the wide range that characterizes the types of AD and the different ways in which they are currently used in the statistical production process, the question arises to what extent it is possible to define generalized methods for their statistical use.

Meanwhile, the NSI's production processes are deeply evolving, not only in the field of business statistics for which the availability of AD is more extensive.

An interesting overview on actual uses of AD for producing business statistics (SBS, STS, Prodcom and Business Register Regulations) in all Member States and EFTA Countries has been conducted by the ESSnet AdminData – Workpackage 1 "Overview of existing practices in the use of AD for producing business statistics over Europe". All information collected is made available to users (internal NSI users) in a Database, which can be browsed by topic, by domain (regulation) and by country. It is possible to get information on the combination of sources used for producing statistics (among survey data, admin data and registers) and on how AD are used: directly for producing statistics, or indirectly for the sampling frame, in editing & validation, in imputation of missing values, in estimation procedures[6].

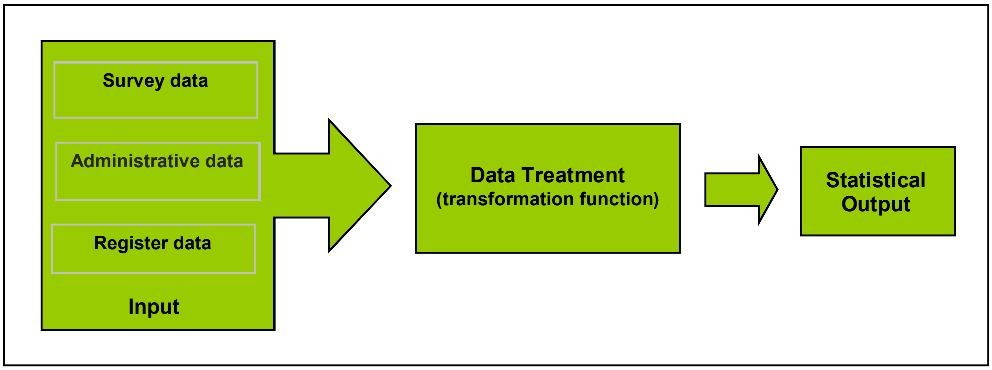
To better focus on the paper objectives, in Figure 1 the statistical process that uses AD is shown. Sometimes the input to the process may consist of a combination of data from different sources and it is also common to use multiple Administrative Sources within the

[5] Nordic countries produced comprehensive documentation of their best practices based on their long experience in using administrative data also for producing censuses data (Unece, 2007). The European Commission programme called MEETS - Modernisation of European Enterprise and Trade Statistics - has funded several activities on this issues in these last years (http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/MEETS_programme); more information for projects that have addressed the AD issue are available at: (a) http://cros-portal.eu/ for ESSnet AdminData (European Statistical System project on The Use Of Administrative And Accounts Data For Business Statistics) had the scope to find common ways for use of dministrative data for business statistics; (b) http://www.blue-ets.istat.it/ for BLUE-ETS project, WP4 that investigated the possibility to increase the use of AD for statistical purposes.

[6] Costanzo et al. (2011), ESSnet AdminData (2013a), Database available on the web site: http://essnet.admindata.eu/.

same process, then data are integrated and treated to produce the statistical output. These are the cases: (a) when an administrative source fails to meet the requirements of quality and multiple sources need to be combined in order to adjust the shortcomings of the separate sources; (b) when data integration produces a much richer information content, such as Linked Employer Employee data or Educational data combined with employment data and so on; (c) for producing longitudinal data (Wallgren and Wallgren, 2007).

Figure 1 - The statistical production process that uses administrative data



Sometimes AD can be used as they are to produce statistical data: this is the limit case where the transformation function in Figure 1 is the identity function. Generally, AD enter the production process directly for producing statistics or registers after the Treatment procedure, or they support the survey process: for producing the sampling frame, in editing & validation process, in imputation of missing values, for unit non response treatment, in estimation procedures[7].

In this paper the problem of defining the quality of AD which enter the statistical production process will be treated. Regardless of the manner in which the AD are used, their quality evaluation in terms of input process is a useful information that has to be associated to AD. In general the lower the quality of the input data and the greater the effort to bring the output data to acceptable quality levels.

The work here presented is based on the results obtained within the BLUE-ETS project, ended in 2013 and specifically Work package 4 (WP4) "Improve the use of administrative sources" aimed to develop an instrument able to determine the statistical quality of AD, a Quality Report Card for Administrative data (QRCA), generally applicable to AD sources in different European countries (Daas et al., 2011a, 2011b, 2013).

The conceptual framework of the QRCA will be presented and relevant quality indicators selected and classified in order to evaluate the statistical usability of AD will be described. In the development of quality indicators associated with the use of AD for statistical purposes, it is useful to distinguish three types of indicators:

[7] It has to be considered that it is not so rare the case in which public administrations directly produce statistical data from their own administrative data as a result of an agreement within the National Statistical System (Sistan in Italy).

1. Input quality indicators: to define the quality of AD used as input in the statistical production process;
2. Process quality indicators: to measure the quality associated with the production process that uses AD to produce statistics;
3. Output quality indicators: to measure the output quality of statistics involving AD, taking input and process quality into account.

In this paper indicators of the type 1 evaluating AD are analyzed. A further useful specification of Input quality indicators, developed within BLUE-ETS WP4 work (Daas et al., 2011), considers the value of the additional information brought to the specific statistical production process by the Administrative Source. 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).

Quality aspects related to the output production will not be considered here. An interesting analysis of the overall quality of register-based statistics is developed in Statistics Sweden (Wallgren, Wallgren, 2007; Laitila et al, 2011). Within the ESSnet AdminData project quality indicators of type 3 have been studied: starting from the state of play in terms of the use of quality indicators for business statistics involving AD across NSIs[8], output quality indicators for statistics involving AD have been produced (ESSnet AdminData, 2013a)[9].

In Section 2 the conceptual framework of the AD quality is described. To test the quality framework, an application to a case-study is reported in Section 3. The Social Security Data source actually in use in more statistical processes and under analysis for other potential uses in Istat has been evaluated as case study. Some general issues on the QRCA usability in Istat are also presented in Section 4.

Before proceeding, it is useful to firstly focus on some definitions used in this paper and some issues related to the context.

The concept of AD quality, as it is here considered, concerns the quality in terms of statistical usability in the production process. For instance an administrative data set with a very good quality for its original purpose may have a poor statistical quality that can affect its statistical usability. Regarding the definitions, we rely on the ESSnet Admin Data Glossary[10] so *Administrative source* and *Administrative dataset* are defined as follows:

Administrative Source

> A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations. In a wider sense, any data source containing information that is not primarily collected for statistical purposes.

Administrative Dataset

> Any organised set of data extracted from one or more Administrative Sources, before any processing or validation by the NSIs.

[8] This work is already in place on the production of Eurostat Quality Report Framework for Business Statistics under Regulation (CE) no. 295/2008 and user test carried out within EU and EFTA countries NSIs.

[9] For more information on ESSnet AdminData - Work Package 6 results see the website http://essnet.admindata.eu/.

[10] ESSnet AdminData Glossary http://essnet.admindata.eu/.

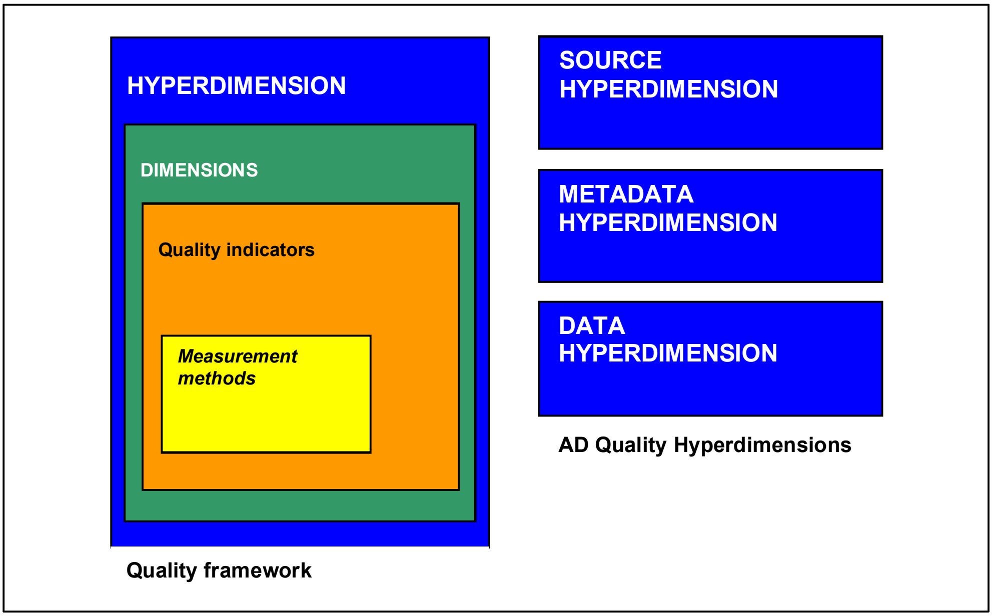
Administrative Data are considered as data derived from an Administrative Source. The term AD quality refers to the statistical quality of the Administrative Dataset provided by AD source holder and received by Istat.

2. The conceptual framework

2.1 A generalized and flexible framework

Following Daas and Ossen (2011), it is useful to define a framework for the statistical quality of AD using a hierarchical and multidimensional approach whose graphical representation is shown in Figure 2. Here the hierarchical aspect is made by the levels in which quality breaks down. Four levels are considered: i) the Hyperdimension (or category) level, ii) the Dimension level that breaks down each Hyperdimension; iii) quality indicators within each Dimension and iv) several measurement methods included in each quality indicator.

Figure 2 - Multidimensional and hierarchical quality framework



Considering secondary data source quality[11], three relevant Hyperdimensions has been selected: (a) the Source itself by considering the data acquisition procedure, the legal basis for its use and a description of the data source use effects on the NSI; (b) the Metadata focused on conceptual aspects such as the units and variables definition, their comparability

[11] "Secondary data source" as defined by Hox and Boeije, 2005: "Data originally collected for a different purpose and reused for another research question". This is a more general definition including also other kind of data such as: commercial data and big data.

with NSI's definitions and the presence of unique identification keys; (c) the Data Hyperdimension related to the data quality (facts).

The multidimensionality is given by the fact that the approach looks at the quality of the AD source as a whole and not only with reference to the Data: that is the reason why the concept of Hyperdimension is introduced.

While the Hyperdimensions, the Dimensions and the Quality indicators are fixed, the Measurement methods are flexible as it is possible to choose within more suggested methods which are the most suitable to measure each indicator. This approach allows for a flexible AD quality evaluation structure capable of defining a generalized tool regardless of:

- the types of AD (tax data, social security data, educational data,…);
- the statistical processes involving AD (business statistics, population statistics, social statistics,…);
- the way AD are used (to directly produce statistics, to form a frame for sample surveys, to provide auxiliary information for sampling design or estimation purpose, for edit and imputation procedures).

As far as the Source and the Metadata Hyperdimensions are concerned, Dimensions and indicators are well described in Daas and Ossen (2011) where a Checklist was developed to assist the preliminary evaluation of AD and support the decision to use or not the source in the statistical production process before acquiring data. The Source and the Metadata quality Dimensions are reported below in Table 1.

Table 1 - AD quality Dimensions and indicators for Source and Metadata Hyperdimension

HYPERDIMENSIONS	Dimensions
Source	1. AD source holder (information for the acquisition data procedure). 2. Relevance of the AD source. 3. Privacy and security. 4. Delivery. 5. Relationships and feedback with the AD source holder.
Metadata	1. Clarity and interpretability . 2. Comparability at the metadata level. 3. Unique keys. 4. Data treatment (by data source keeper).

In the next Section the Data Hyperdimension components, as defined in the BLUE-ETS WP4 project, will be presented. Results from the application of quality indicators on AD should be collected and shown in a quality report, the QRCA, which is an instrument for the input quality evaluation in statistical production processes based on AD.

Obviously the indicators presented are producer statistics-oriented and not user statistics-oriented since their aim is to support the statistical production within NSIs. In paragraph 4 the objectives of the QRCA are described considering the different type of users.

2.2 Dimensions, indicators and measurement methods

The topic of this section is to firstly define the Dimensions selected to evaluate the quality related to the Data Hyperdimension and secondly to describe the selected indicators within the Dimensions.

Measuring AD statistical quality is not the same thing as measuring statistical survey data quality. It is well known that the quality of statistics (i.e. the statistical output) is defined by Eurostat with reference to the following six quality Dimensions (Eurostat 2003a, 2003b, 2005):

- relevance;
- accuracy;
- timeliness and punctuality;
- accessibility and clarity;
- comparability;
- coherence.

However these criteria do not apply to AD as they are not able to correctly or exhaustively evaluate all the quality aspects of AD considered as the input of the statistics production process. Therefore a further effort has to be made to revise the existing statistical survey quality framework and to formalize a new one useful for the AD quality evaluation.

The five quality Dimensions selected for AD by the BLUE-ETS WP4 project are described and reported below in Table 2. From a general point of view they focus on the AD quality from the moment they arrive in the NSI, till the moment they are available as input for the statistical production process. Having said this, Eurostat components formerly defined to report on the quality of survey statistics (users statistics-oriented) should be totally reviewed into an input quality perspective (producer statistics-oriented). Hence this means a change of the viewpoint for some existent Dimensions, like Accuracy and Completeness, and the introduction of new Dimensions aimed to evaluate some very crucial aspects of the AD quality, like Technical checks and Integrability, that are totally new with respect to the survey's world as described below (Table 2). In particular the new Integrability Dimension is of primary importance since AD concepts, rules and classification criteria for objects and variables are different from the NSI's ones and to make AD usable for statistical purposes often a reconciliation has to be made (Wallgren and Wallgren, 2007). Integrability Dimension measures the extent to which AD can be integrated in the statistical production process.

Table 2 - AD quality Dimensions in the Data Hyperdimension

DIMENSION	Description
Technical checks	Technical usability of the file and data in the file.
Integrability	Extent to which the data source is capable of undergoing integration or of being integrated.
Accuracy	The extent to which data are correct, reliable and certified.
Completeness	Degree to which a data source includes data describing the corresponding set of real-world objects and variables.
Time-related Dimension	Indicators that are time or stability related.

The AD quality Dimensions may be applied to different levels: the entire dataset, the objects[12] and the variables. Actually this classification allows maintaining simple, comprehensive and compact the theoretical structure. Quality indicators by Dimension and level of application are shown in Table 3.

Table 3 - Quality indicators by Dimension

INDICATORS BY DIMENSION	Level	Description
1. Technical checks		
Readability	Dataset	Accessibility of the file and data in the file.
Convertibility	Objects	Conversion of the file to the NSI-standard format.
File declaration compliance	Variables	Compliance of the data in the file to the metadata agreements.
2. Integrability		
Comparability of objects	Objects	Similarity of objects in source with the objects used by NSI.
Alignment of objects	Objects	Linking-ability (align-ability) of objects in source with those of NSI.
Linking variable	Variables	Usefulness of linking variables (keys) in source.
Comparability of variables	Variables	Proximity (closeness) of variables.
3. Accuracy		
Authenticity	Objects	Legitimacy of objects.
Inconsistent objects	Objects	Extent of erroneous objects.
Dubious objects	Objects	Presence of untrustworthy objects.
Measurement error	Variables	Correctness of a value with respect to the measurement process.
Inconsistent values	Variables	Extent of inconsistent (out of range) variable’s values or combinations of values for variables.
Dubious values	Variables	Presence of implausible values or combinations of values for variables.
4. Completeness		
Under-coverage	Objects	Absence of target objects (missing objects) in the dataset.
Over-coverage	Objects	Presence of non-target objects in the dataset.
Selectivity	Objects	Statistical coverage and representativeness of objects.
Redundancy	Objects	Presence of multiple registrations of objects.
Missing values	Variables	Absence of values for (key) variables.
Imputed values	Variables	Presence of values resulting from imputation actions by data source holder.
5. Time-related Dimension		
Timeliness	Dataset	Lapse of time between the end of the reference period and the moment of receipt of the dataset.
Punctuality	Dataset	Possible time lag between the actual delivery date of the dataset and the date it should have been delivered.
Overall time lag	Dataset	Overall time difference between the end of the reference period in the dataset and the moment the NSI has concluded that it can definitely be used
Delay	Dataset	Extent of delays in registration.
Dynamics of objects	Objects	Changes in the population of objects (new and dead objects) over time.
Stability of variables	Variables	Changes of variables or values over time.

[12] The term “object” is used to generalize and to include events registered or units of a target set (administrative population); from objects should be possible to derive statistical units (Zhang, 2011).

For example, within the Accuracy Dimension inconsistency is evaluated for objects as well as for variables in order to have a complete view of the extent to which data are correct, reliable and certified. Same symmetry occurs for Completeness too as the assessment can take place in terms of objects (under and over coverage indicators) and in terms of variables (indicators for missing values).

To draw a parallel and alight the differences with the survey data quality assessment (Eurostat, 2003b), it has to be said that the Dimensions of Accuracy and Completeness are fully reviewed. In surveys under and over-coverage errors are indicators included into the Accuracy Dimension as well as the missing values that are “item non response”. In AD quality the coverage is, instead, evaluated only *a posterior* with respect to the statistical target population and therefore it is not connected to the Accuracy Dimension but concerns the data completeness. This approach is typical of AD because the statistical population is not defined *a priori* as in the case of statistical surveys.

Missing values are included into the Completeness Dimension too as they may arise from inaccuracy of the source, but also by the fact that most of the times some variables are not mandatory for subpopulations while they are considered relevant for statistical purposes: so it is a problem of completeness rather than accuracy.

It is important to note that the quality Dimensions are not mutually orthogonal and, as a matter of fact, some trade - offs are present. For instance it often happens that, to get timely data, a lower quality in the Completeness Dimension is considered acceptable, for example for the production of short-term statistics.

From the operational point of view, several R scripts have been developed that allow to automatically calculate some suggested indicators and provide attractive graphical displays of data that can help the application of the quality framework within the NISs (Tennekes et al., 2011).

After a general overview on the AD quality Dimensions, in the following paragraphs indicators and measurement methods are described.

2.2.1 Technical checks and indicators

The Technical checks Dimension includes the IT-related indicators for the data in a source through which the technical usability of the file and data in the file are verified (Daas et al., 2011b). This Dimension includes three indicators:

1. the Readability;
2. the Convertibility;
3. the File declaration compliance.

All these indicators are associated to the dataset and check respectively whether (1) it is impossible to physically access the data as the file cannot be opened or it is damaged, or (2) data are not correctly convertible into NSI-standard formats, or (3) the data delivered are not conform to the definitions included in the metadata, if any are provided, or data do not comply with the request of the NSI.

Since this Dimension looks at the technical usability of files and data, it is useful to remark that the corresponding quality indicators are important:

a) to support the data loading and decide whether carrying on using the data source or going back to the ADH because of errors in the delivery;
b) for monitoring data deliveries when the source becomes in use fully operative;

c) to provide assistance when a new data source is being studied and its technical usability is explored for the first time.

To have an overall look of the data provided, with the aim to identify obvious errors in supply, it is helpful to use the Tableplot function available in the R package (Tennekes et al., 2011).

Some measurement methods for the Technical Checks indicators are proposed in the following Table 4.

Table 4 - Suggested measurement methods by indicator for Technical checks

INDICATORS BY DIMENSION	Measurement methods
Readability	a. % of deliveries (or files) of the total deliveries with an unknown extension, that are corrupted or cannot be opened b. % of the total file which is unreadable (MB/GB size) or number of unreadable records
Convertibility	% of objects with decoding errors or corrupted data.
File declaration compliance	% of variables in the current delivery that differ from metadata lay-out delivered or agreed upon in: i) formats and names ii) variable and attribute content iii) categories defined for categorical variables iv) ranges for numerical variables

2.2.2 Integrability and indicators

This Dimension contains indicators aimed to evaluate the ease by which the data in the source can be integrated into the statistical production system of an NSI. Since it immediately looks at the integration process, the idea standing behind this Dimension is clearly an IOQ view (Daas et al., 2011b).

It should be noted that the Integrability is a characteristic Dimension in assessing the input quality of an AD source. In facts, since AD are primarily collected for non-statistical purposes and describe non statistical concepts (e.g. administrative, fiscal) they firstly need to be converted into statistical concepts by appropriate harmonization. The reconciliation of concepts and definitions for objects and variables on AD source plays an important role for the source evaluation in terms of the Integrability Dimension before data are actually integrated and involves both Metadata and Data Hyperdimensions.

The Integrability Dimension includes indicators for objects and for variables.

Comparability and Alignment indicators evaluate the similarity of objects in the Administrative Source and their linking-ability with those used in the statistical production system by measuring the distance from the point of view of the objects definition. Administrative objects are analyzed with respect to their degree of comparability with the statistical objects and they are evaluated as identical, corresponding or incomparable objects according respectively to the fact that (a) they have exactly the same unit of analysis and definition as those used by NSI or (b) they correspond after harmonization or (c) they are not comparable.

Some measurement methods for indicators are proposed in Table 5.

Table 5 - Proposed measurement methods for Integrability indicators

INDICATORS BY DIMENSION	Measurement methods
Comparability of objects	a. % 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 b. % 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 c. % 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 d. % 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	a. % of identical aligned objects = (Number of objects in the reference statistical population with exactly the same unit of analysis and same concept definition as those in the source) / (Total number of relevant objects in the reference statistical population) x 100 b. % of corresponding aligned objects = (Number of objects in the reference statistical population that, after harmonization, correspond to units or parts of units in the source) / (Total number of relevant objects in the reference statistical population) x 100 c. % of non-aligned objects = (Number of objects in the reference statistical population that, even after harmonization of the objects in the source, cannot be aligned to one of the units in the source) / (Total number of relevant objects in the reference statistical population) x 100 d. % of non-aligned aggregated objects = (Fraction of objects of interest at an aggregated level in source 1 that cannot be aligned + fraction of objects of interest at the same aggregated level in source 2 that cannot be aligned) x 100
Linking variable	a. % of objects with no linking variable = (Number of objects in source without a linking variable) / (Total number of objects in the source) x 100 b. % of objects with linking variables different from the ones used by NSI = (Number of objects in source with linking variables different from the one used by the NSI) / (Total number of objects with linking variables in the source) x 100 c. % of objects with correctly convertible linking variable = (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	a. 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 b. The Mean Absolute Percentage Error (MAPE) that measures the mean of the absolute percentage error c. A method derived from the chi-square test that evaluates the distributions of the numeric values in both data sets. For categorical data Cramer's V (Cramer, 1946) could be used. The comparison could be performed either by groups (macro level) or at the micro level d. % 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

All these methods highlight the complexity of comparing sources and how important is the Integrability Dimension for evaluating the statistical usability of an AD source and its integration in the statistical process.

2.2.3 Accuracy and indicators

Indicators in this Dimension are derived from the sources of error scheme for AD firstly proposed by Bakker (2010) and developed by Zhang (2012) where errors occurring during the collection of AD are described from the moment they are collected by the ADH up to the moment when the data are linked to other statistical data sources to be used as input for NSI. Accuracy of AD is defined as the extent to which data are correct, reliable and certified. Some measurement methods for indicators are proposed for objects and variables in the following Table 6.

The indicators for objects evaluate the correctness of the units or events registered in the source and include the Authenticity which measures the objects correspondence to the real world and the Inconsistent or dubious objects indicators that check for the objects involvement in respectively inconsistent or dubious relations.

The indicators for variables evaluate the validity of the variables as correctness of the values of the units or events registered in the source. These indicators include the Measurement errors which measure the correctness in terms of deviation of data value from ideal error-free measurements and the Inconsistent or dubious values that check for objects with values involved in respectively inconsistent or dubious relations.

It is necessary to highlight the difference with the concept of Measurement errors used in the data survey quality. The Measurement errors evaluate the distance between the observed value and the true value but in the administrative source they are by definition out of the NSI control as these errors are the result of the data collection carried out by the ADH. Thus the measurement errors could only be known ex post in an indirect way by asking the ADH information about the quality checks management during the data collection phase if any exists.

Table 6 - Proposed measurement methods for Accuracy indicators

INDICATORS BY DIMENSION	Measurement methods
Authenticity	a. % of objects with a non-syntactically correct identification key b. % of objects for which the data source contains information contradictive to information in a reference list for those objects (master list and target list) c. Contact the data source holder for their % of non-authentic objects in the source
Inconsistent objects	% of objects involved in non-logical relations with other (aggregates of) objects.
Dubious objects	% of objects involved in implausible but not necessarily incorrect relations with other (aggregates of) objects.
Measurement error	a. % of unmarked values in the data source for each variable (when values not containing measurement errors are marked by AD holder) b. 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?
Inconsistent values	% of objects with inconsistent (out of range) variable's values or objects whose combinations of values for variables are involved in non-logical relations.
Dubious values	% of objects with dubious variable's values or objects whose combinations of values for variables are involved in implausible but not necessarily incorrect relations.

2.2.4 Completeness and indicators

This Dimension is defined as the degree to which a data source includes data describing the corresponding set of real world objects and variables. Thus the indicators for objects in this Dimension mainly focus on coverage topics while the indicators for variables are related to missing and imputed values.

Actually the indicators for objects are the indicators of (1) Under-coverage, (2) Over-coverage, which measures the coverage with respect to a target statistical population, (3) the Selectivity which evaluates the coverage by specific statistical subpopulations and (4) the Redundancy that checks for duplications in the recording of objects.

The indicators for variables are indeed the Missing values - that evaluate objects with completely or partially missing values for key variables (missing units and missing items) and Imputed values that calculate objects with values imputed by the ADH.

It is useful to highlight that the way to calculate indicators of Under-coverage, Over-coverage and Selectivity is twofold. In facts they can be computed with respect to the statistical target populations (that means an IOQ view) or to the administrative target population of the source (in a DSQ view). It should be noted that most of the times the statistical target population is not available for timeliness reasons while the administrative target population generally may not exist at all.

As far as Redundancy is concerned, although it measures the quality of the delivered data by counting multiple registrations of data objects in the source, however it is important to underline that sometimes same object may have multiple registrations because it is a characteristic of the AD source (e.g. in more registrations the AD source records several information about the same employee).

Some measurement methods for indicators are proposed in the following Table 7.

Table 7 - Proposed measurement methods for Completeness indicators

INDICATORS BY DIMENSION	Measurement methods
Under-coverage	% of objects of the reference list missing in the source.
Over-coverage	a. % of objects in the source not included in the reference population b. % of objects in the source not belonging to the target population of the NSI
Selectivity	a. 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 b. Use of more advanced graphical methods, such as table plots c. Calculate the Representativeness indicator (R-indicator; Schouten et al, 2009) for the objects in the source
Redundancy	a. % of duplicate objects in the source (with the same identification number) b. % of duplicate objects in the source with the same values for a selection of variables c. % of duplicate objects in the source with the same values for all variables
Missing values	a. % of objects with a missing value for a particular variable b. % of objects with all values missing for a selected (limited) number of variables c. Use of graphical methods to inspect for missing values for variables
Imputed values	a. % of imputed values per variable in the source b. Contact the data source holder and request the percentage of imputed values per variable

2.2.5 Time-related Dimension and indicators

The quality indicators in this Dimension are all related to time. The Timeliness, the Punctuality, and the Overall time lag indicators apply to the delivery of the administrative dataset. The Overall time lag indicator, which measures the total time lag between the reference period and the moment at which data can be used by the NSI, also includes the time required for evaluation. The Delay indicator is built up to evaluate how fresh is the information stored in the AD source with respect to the real world events as it aims to measure the extent of delays in registration.

These indicators are all relevant for both DSQ and IOQ.

The last two indicators apply to objects (Dynamics of objects) and to variables (Stability of variables). The indicator for objects aims to describe changes in the population over time by comparing the population of objects referred to time t (delivery τ) to that referred to time t-1 (delivery $\tau-1$). In this case the indicator does not express a direct evaluation of the data quality (the population dynamics depends on the phenomenon under study), but only a description of the dynamics. The indicator of Stability of variables describes the stability in terms of the changes over time in variable values on persistent objects at time t (delivery τ) compared to those at time t-1 (delivery $\tau-1$). Here the attention is focused on the variable composition covered by a source that should be stable in time between subsequent deliveries (e.g. the company Nace code).

It should be noted that concerning data over time it is very useful to consider tools for analysis of time series and longitudinal data.

Some measurement methods for indicators are proposed in the following Table 8.

Table 8 - Proposed measurement methods for Time-related Dimension indicators

INDICATORS BY DIMENSION	Measurement methods
Timeliness	a. Time difference (days) = (Date of receipt by NSI) – (Date of the end of the reference period over which the data source reports) b. 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	a. Contact the data source holder to provide their information on registration delays b. Time difference (days) = (Date of capturing the change in the source by the data source holder) – (Date the change occurred in the population)

Table 8 continued - **Proposed measurement methods for Time-related Dimension indicators**

INDICATORS BY DIMENSION	Measurement methods
Dynamics of objects	a. % Births t = (Births t / Total objects t) x 100 = (Births t / (Births t + Alive t) x 100 b. % Deaths t = (Deaths t / Total objects t) x 100 = (Deaths t /(Births t + Alive t) x 100 c. % Deaths t-1 = (Deaths t / Total objects t-1) x 100 = (Deaths t / (Alive t + Deaths t) x 100
Stability of variables	a. 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 b. % of Changes = (Number of objects with a changed value / total number of persistent objects with a value filled in for the variable under study) x 100 c. 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 Cramer's V can be used

3. Application to a case-study: the SSD administrative source

3.1 Description of the source used for the case-study

Italian Social Security Data (SSD), used for this application of the input quality indicators, is produced by Inps (Italian Institute of Social Security) and concerns the monthly contribution declarations of employers for employees, as requested by the law of 24 November 2003, n. 326. The so-called Emens declarations are the means by which Inps collects pay data and information needed to computation of the social security contributions for each employee.

The choice of the SSD source for testing QRCA quality framework derives from several elements: it is a complex and big source including more statistical units connected to each other; due to its wealth of information it is suitable to be widely used in Istat within different statistical processes, both for the production of business and social statistics.

Particularly relevant is the central role that the SSD source has assumed in redefining the Istat production process of the Business Register (BR), Asia (Archivio statistico delle imprese attive) for the 2011 edition and in implementing innovative information about employment for the Census of Industry and Services 2011, mainly based on AD.

The SSD source covers data on the social security system for private employers as well as for other small subsets of public employers. Regarding the territorial reference, SSD includes social security contributions payable by employers resident in Italy.

Until 2010, Inps provided to Istat an annual dataset built on the basis of the monthly declarations and data referred to year *t* were delivered after 18 months.

From the supply of 2010 onwards, except for some previously test, the Istat interest has moved to the monthly version. Monthly data referred to year *t* are provided in two releases, on April *t+1* and on November *t+1* with an improvement of the timeliness (a maximum of 12 months for the final version). Data supply is very big, about 160 million records and 45 variables. For the application of the quality framework, measurement methods of quality indicators are computed on May 2010 data. This subset of the entire database contains about 13 million records and the same number of variables.

As already mentioned, more types of objects can be identified in SSD. Each record refers to the “Employee Tax Feature” defined as the set of variables useful to calculate 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 Employee Tax Feature are the Type of employment contract (Fixed term/Permanent), the Contractual working time (Full/Part-time), the Professional status and the Type of contribution (tax relief for disabled or disadvantaged workers,…). 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.

Some units that can be derived from the Employee Tax Feature are: the Employee, the Employer and the Workplace (Municipality).

In Table 9, the administrative units are described and their identification key variable is reported. The Employee and Employer units are both identified through the Tax Code; concerning the Workplace, a code called "Belfiore" is used to identify the Italian Municipalities.

Table 9 - Administrative units in the SSD source

DEFINITION	Identification key
EMPLOYEE TAX FEATURE – primary unit The set of characteristics useful to define the amount of social security contributions payable for each employee by the employer	Complex key defined by a set of variables.
EMPLOYEE – derived unit A worker who has had at least one pay contributions to INPS as an employee during the month	Employee Tax Code.
EMPLOYER – derived unit 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	Employer Tax Code.
WORKPLACE – derived unit Place where the work is mainly carried out	Belfiore Municipality Code.

Among the SSD administrative units there are several relationships:

a) an Employer may have more Employees;
b) an Employee can have more than one Tax Feature with the same Employer;
c) an Employee can have more than one Tax Feature with the more than one Employer;
d) a Municipality can be host for more Employees;
e) (a Municipality may not be a Workplace, in this case it is not recorded in the dataset).

The SSD source provides a wealth of information useful to describe the employment in enterprises. In addition to the main variables already mentioned for describing Employee Tax Feature, it contains 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 defined in the monthly data with the day of the month.

3.2 Working method and selection of the indicators

The results presented here on the SSD derive from the methodological application carried out by Istat that has been called, along with Statistics Netherlands and Statistics Sweden, to test the theoretical framework defined to evaluate AD quality within the BLUE-ETS WP8 (Daas et al., 2013).

The same results, however, were here taken up and examined in the attempt to respond to a different purpose, which is to determine the applicability of the QRCA into the Istat production processes, at the Data Hyperdimension level. We are not considering here the Source and the Metadata Hyperdimensions involving the AD acquisition task and Metadata analysis task already mentioned (Daas and Ossen, 2011).

The feasibility study for the production of the QRCA for SSD means to identify, first of all, which indicators and measurement methods included in the framework BLUE-ETS are actually applicable, useful and computable on each supply of SSD source, as soon as it is acquired by Istat. The aim, in this case, is to identify the set of information on the SSD quality that is possible to release in a timely manner in the moment in which the data are delivered to the Istat users of the source. Timeliness is a central element on which to focus, in particular for sources already included in the statistical processes, as is the case of the SSD. For these sources, in fact, the time that elapses between the acquisition (by the directorate in charge) and the delivery of data to users can be very short, and it is in this time that the BLUE-ETS indicators should be calculated.

The implementation of the quality indicators depends on two main aspects: the availability of the information requested for the measurement methods computation and the possible automatic IT procedures that should make the computation as timely as possible.

With respect to the second aspect, the activities are in progress in Istat.

This paper focuses instead on the first aspect and presents a preliminary analysis aimed at assessing the applicability of the indicators and related measurement methods verifying which information requested for the computation is actually available about SSD. In order to verify this, quality indicators have been divided between those referring to the entire dataset (Technical checks and some Time related Dimension indicators), and those applicable to selected units or variables (Integrability, Accuracy, Completeness and some Time related Dimension indicators). For the second group of indicators it was necessary to perform some preliminary actions:

- selection of objects in the SSD source to which it was possible and useful to calculate indicators;
- identification and acquisition of the reference statistical population to compare and match data (Integrability and Completeness Dimensions);
- selection of variables in the SSD source to which it was possible and useful to calculate indicators;
- identification of relevant edit rules to which it was possible and useful to calculate indicators (Accuracy Dimension).

The analysis carried out in the test phase (Daas et al., 2013) had already pointed out that problems of applicability do not occur for indicators whose calculation depends on information included directly in the dataset provided, while some issues emerge when the information requested is external to the supply and are primarily due to:

- the not comprehensive information given by the AD provider on the metadata and on the data generation process;
- the existence of the reference list (administrative target population or statistical target population) for the comparison of the units or variables.

Regarding the last point, if in the test case indicators of Comparability, Alignment, Undercoverage and Overcoverage were computed for Enterprise unit, using the BR as the reference statistical population, in this context they have been excluded for two reasons: one connected to timeliness issues and the other one to a more substantial restriction.

The supply of SSD used in this work was acquired on November 2011, while Asia register related to 2010 was made available on January 2012, in draft form, and on May 2012, in the final version. It is evident that the timeliness requirement for the computation of the indicators is not respected.

About the second reason, innovations introduced from 2011 in the Asia BR production made the SSD source part of the input sources used. This dependence relationship makes that it will never be possible to calculate the indicators mentioned using the Asia register as reference statistical population.

For Comparability and Alignment indicators it is essentially a practical problem and an approximate solution may be to use the BR of the previous year to evaluate the Integrability. The calculation was not carried out in this work. However, an analysis of Integrability was performed at the Metadata level on the Employer unit in the SSD source and the Enterprise unit in the Asia register (see Table 9 in the Section 3.3). Useful information about the Integrability of the SSD may be obtained by calculating the indicators also with respect to other AD sources, such as those that enter along with it in the Asia production process.

In the case of the Undercoverage and Overcoverage indicators, the dependence of the Asia register on SSD source poses, instead, a problem that is, not only practical, but also, and especially, of a conceptual nature. The independence between the AD source under evaluation and the reference statistical population should be an essential requirement when analyzing the coverage of an AD source in terms of input quality.

Table 10 lists the set of input quality indicators that can be calculated on the basis of the analysis carried out comparing the information requested for the computation with those actually available. For indicators related to objects, the type of object on which the indicator has been applied is reported. It is pointed out that whereas in SSD objects are repeated on multiple records (by construction), the indicators were calculated by properly counting the number of records or the number of units as specified.

Table 10 - Indicators applied to SSD

DIMENSION	Level	Indicator	Type of object			
			Employee	Employer	Municipality	Employee Tax Feature
Integrability	Object	Comparability of objects			x	
	Object	Alignment of objects			x	
	Variable	Linking variable				
Accuracy	Object	Authenticity	x	x		
	Object	Dubious objects	x			
	Variable	Inconsistent values				
	Variable	Dubious values				
Completeness	Object	Redundancy	x	x	x	x
	Variable	Missing values				
Time related Dimension	Dataset	Timeliness				
	Dataset	Punctuality				
	Objects	Dynamics of objects	x	x	x	

In the next Section results of the indicators related to the SSD are reported. For the implementation of the measurement 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 it is the case of SSD.

3.3 Case-study results

In this Section results of the quality indicators applied on the SSD administrative source are presented by quality Dimension.

Starting with the Integrability Dimension, we focused on Alignment and Comparability indicators, for objects, and on Linking variables indicator. With regard to the two indicators of objects a preliminary analysis at the Metadata Hyperdimension level is required and it is made comparing administrative concepts, already described (Table 9), with the statistical ones, as shown in the following Table 11. For the Employee and Employer units there is a similarity (*identical objects*) between administrative and statistical concepts and SSD identification variables are the same used by Istat (Tax Code). The units Employee Tax Feature and Municipality are not directly comparable with statistical units of interest. But, after a treatment process, it is possible to integrate them (*corresponding objects*). In particular, the Employee Tax Feature 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 Workplace it is possible to use an external table to harmonize administrative and statistical units.

Table 11 - Administrative units in the SSD source and Reference statistical units used in Istat

ADMINISTRATIVE OBJECTS		INTEGRABILITY LEVEL	STATISTICAL OBJECTS	
Definition	Identification key		Definition	Identification key
EMPLOYEE TAX FEATURE The set of characteristics useful to define the amount of social security contributions payable to INPS for each employee by the employer	Complex key defined by a set of variables	Corresponding ⟺	EMPLOYMENT RELATIONSHIP A formal agreement between an enterprise and a person, whereby the person works for the enterprise in return for remuneration	- (Employment relationships register not available)
EMPLOYEE – derived unit A worker for which there is at least one social security contribution paid to INPS as an employee during the month	Employee Tax Code	Identical ⟺	EMPLOYEE IN ENTERPRISE Person who works for an Enterprise on the basis of a contract of employment and receives compensation	- (Employees register not available)
EMPLOYER – derived unit Employer who has employed at least one regular worker	Employer Tax Code	Identical ⟺	ENTERPRISE Enterprise in Business Register ENTERPRISE WITH EMPLOYEES Enterprise with employment >0 in Business Register	Enterprise Tax Code
WORKPLACE - derived unit Place where the work is mainly carried out	Belfiore Municipality Code	Corresponding ⟺	Italian Municipality	Istat Municipality Code

As explained in Section 3.2, in the Data Hyperdimension, the Comparability and Alignment indicators for Employers are not reported, as the BR cannot be used in practice for the computation.

It is however possible to calculate the two indicators for the Workplace (Municipality), using the official Istat Municipalities Register[13]as reference statistical population. The list used is the one updated on the 1st of January 2011 when the official number of Italian municipalities was to 8,094 units.

As already said, the Municipality unit in SSD is not directly comparable with the statistical unit in Istat Register so it is classified as "Corresponding" (Table 11). However a table is available for the harmonization between the Municipality identification codes (named Belfiore Code) in SSD and the Istat Municipality Codes (foreign key). Administrative units are involved in (n : 1) with (n ≥1) relations with statistical units.

Comparability and Alignment indicators are presented in the following Table 12.

Table 12 - Comparability and Alignment of objects indicators

INDICATOR	Measurement method	Result
Comparability of objects	% of the SSD Municipalities corresponding to Istat Municipalities.	97,49%
Alignment of objects	% of the Istat Municipalities corresponding to SSD Municipalities.	99,62%

[13] The register 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.

The results express the weight of the similarity between the two sources. In the Comparability indicator, this weight is calculated with respect to SSD, while in the Alignment indicator with respect to the Istat Register of Italian Municipalities. Users can draw their own conclusions: the degree of Comparability is good and the decoding table works fine (less than 3% of the Municipalities in SSD are not found in the Istat Register); with regard to the Alignment, it is possible to conclude that information is exhaustive as only few Istat Municipalities are not corresponding due to the fact that not all the Municipalities are workplaces (the comparison would have been perfect using the hypothetical Register of the Municipality that are places of employment).

The Linking variables indicator, reported in Table 13, gives information about the usability of units identification codes in SSD for integration with other micro data sources. It has been computed for Employee Tax Code, Employer Tax Code and Belfiore Municipality Code. The results show a high quality of the linkage variables.

Table 13 - Linking variable indicator

INDICATOR	Linking variable	Measurement method	Result
Linking variable	Employee Tax Code	% of records in SSD with missing value on the Employee Tax Code.	0,00015%
		% of records in SSD with syntactical incorrect value on the Employee Tax Code.	0%
	Employer Tax Code	% of records in SSD with missing value on the Employer Tax Code.	0%
		% of records in SSD with syntactical incorrect value on the Employer Tax Code.	0%
	Belfiore Municipality Code	% of records in SSD with missing value on the Belfiore Municipality Code.	0,06%
		% of records in SSD with Belfiore Municipality Code convertible to one used by Istat.	99,89%
		% of Municipalities in SSD with Belfiore Municipality Code convertible to one used by Istat.	97,49%

The indicators of the Accuracy Dimension allow to provide an assessment of the correctness of SSD both for objects and for variables. With regard to the objects, the Authenticity and the Dubious objects indicators are shown in the following Table 14. The first focuses on the legitimacy of objects and it is calculated for the Employee and Employer units checking the syntactic correctness of their identification codes. The result coincides with that previously reported for Linking variable indicator (see Table 13). This is an example of how the same measurement method can be used to evaluate two different aspects of the AD quality.

Table 14 - Accuracy of objects indicators

INDICATOR	Object	Measurement method	Result
Authenticity	Employee	% of records in SSD with syntactical incorrect value on the Employee Identification Code (Tax Code).	0%
	Employer	% of records in SSD with syntactical incorrect value on the Employer Identification Code.	0%
Dubious objects	*Employee (in relation to Employer)*	% of Employees in SSD which worked at more than 4 Employers.	0,0099%

Dubious objects indicator can be measured investigating the correctness of each object with respect to other types of objects in SSD. A soft rule can be defined to detect objects involved in implausible but not necessarily incorrect relations. In this application, we investigated the relation between the Employee unit and the Employer unit counting the number of attachments for each Employee with different Employers registered on May 2010. The distribution is reported in Table 15.

Table 15 - Distribution of Employees by number of Employers

EMPLOYERS	Employees
1	12.712.004
2	269.496
3	14.708
4	2.506
>= 5	1.283
Total	**12.999.997**

For the calculation of the indicator in Table 14, we considered the following soft rule for each Employee: *More than k "attachments" with different Employers during the month.* The value of the parameter k could be, for example, k = 4. The Dubious objects indicator provides the percentage of units that should 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.

Concerning Accuracy of variables in SSD, we focused on the Inconsistent values and Dubious values indicators. For the calculation, a set of checking rules – respectively, hard and soft rules – should be defined and applied to the variables in the dataset. The rules here examined are to be considered only by way of exercise. The overall definition of these rules requires a thorough knowledge of data and therefore the involvement of the Istat researchers using SSD. Table 16 shows some results of the two indicators applied to the dataset, reporting the percentage of records for which each rule is violated.

Table 16 - Accuracy of variables indicators

INDICATOR	Measurement method	Rule	Result
Inconsistent values	% of records in SSD of which values (or combination of values) for variables are involved in non-logical relations.	*Hard rule* Full-time employment and zero part-time percentage.	0,11%
Dubious values	% of records in SSD of which values (or combination of values) for variables are involved implausible but not necessarily incorrect relations.	*Soft rules* Employee age <= 65.	0,37%

Focusing on the Completeness Dimension, on the basis of available information in SSD, Redundancy and Missing value indicators are the only indicators for which it is possible to meet the timeliness and independence criteria adopted.

The Redundancy has been measured for the different types of objects detecting duplicates for the respective identification codes. For the Employee Tax Feature unit, a multiple identification code is assumed considering the following set of variables:

Employee Tax Code, Employer Tax Code, Professional status, Contractual working time, Type of employment contract, Type of contribution. A last Redundancy indicator has been calculated also to check the occurrence of multiple records, with the same values for all variables. As shown in Table 17, in SSD there are no duplicated records for the entire set of variables, while high percentages of duplicates are found for objects. It should be noted that while the presence of duplicated records with the same values for all variables has to be evaluated as an error and detects problems of data quality, the presence of duplicates on the identification codes for objects is admissible and it depends on the mechanism of data generation (see § 3.1).

Table 17 - Redundancy indicator

INDICATOR	Object	Measurement method	Result
Redundancy	Employee	% of records in SSD duplicated for Employee Tax Code.	2,81%
	Employer	% of records in SSD duplicated for Employer Tax Code.	88,72%
	Municipality	% of records in SSD duplicated for Municipality Belfiore Code.	99,93%
	Employee Tax Feature	% of records in SSD duplicated for Employee Tax Feature multiple identification code.	0,47%
	-	% of records in SSD duplicated for all variables.	0%

Regarding the Completeness of variables, Missing values indicator has been calculated counting the number of records with missing values for the main SSD variables. This indicator can be implemented easily and in a timely manner, as it requires no additional information other than that contained in the dataset itself. As reported in Table 18, the percentage of missing values is equal, or very close, to zero for all the variables considered. 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. In some cases, the latter can be identified in association with the value of the corresponding possible filter variable. In SSD, this happens for the variables: Hiring reason, Job contract termination reason and Part-time percentage. In particular, for the Hiring reason and the Job contract termination reason a value is expected if the respective date is "active" while for the Part-time percentage a value is expected if the Contractual working time is equal to 'Part-time'. It can be useful to evaluate the presence of missing values considering more variables simultaneously. In this application, we calculated the percentage of records with all missing values for the set of variables, already considered in the Redundancy indicator, that it is expected to be the multiple identification code of the Employee Tax Feature.

A graphical representation of the number of missing values may be useful in preparing the quality report.

Table 18 - Missing values indicator

INDICATOR	Measurement method	Variable	Result
Missing values	% of records in SSD with missing value for a particular variable.	Professional status.	0%
		Contractual working time.	0,03%
		Type of employment contract.	0,03%
		Type of contribution.	0%
		Hiring date.	0%
		Hiring reason.	0%
		Job contract termination date.	0%
		Job contract termination reason.	0%
		Part-time percentage.	0%
	% of records in SSD with all missing for a set of variables.	Employee Tax Feature multiple identification code.	0%

With regard to the Time-related Dimension, in this application we considered the Timeliness and Punctuality indicators, referred to entire dataset, and the Dynamics of objects indicator.

Timeliness and Punctuality indicators have been calculated with the aim to measure, respectively:

- the time difference (days) between the date of receipt by Istat and the end of the reference period.
- the time difference (days) between the date of receipt by Istat and the date of receipt agreed upon, as defined in the agreement with the AD provider.

Results of both indicators (Table 19) point out a good quality of SSD and their possible use in the statistical production processes in a timely manner. The Punctuality indicator, assuming a negative value, shows that data were delivered before the receipt date specified in the official request.

Table 19 - Timeliness and Punctuality indicators

INDICATOR	Measurement method	Result (days)
Timeliness	Time difference (days) between the date of receipt of SSD by Istat and the end of the reference period.	365
Punctuality	Time difference (days) between the date of receipt of SSDby Istat and the date of receipt agreed upon, as laid down in the contract.	-31

The Dynamics of objects indicator gives information about changes over time of the populations present in SSD. As the dataset contains monthly data, it was considered useful to point out the dynamics between two consecutive months of the same provision, comparing objects in the months of April 2010 (t-1) and May 2010 (t). In Table 20, results are provided both for Employers and for Employees, in a longitudinal perspective, performing a microdata record linkage between the two monthly datasets. For the Employees, the Dynamics indicator is equal to 3% if we consider the “new workers” and to 2.3% referring to “old workers”. For Employers, values are lower (2.5% and 1.8%, respectively) showing a more limited dynamics.

Table 20 - Dynamics of objects indicator

INDICATOR	Object	Measurement method	Result
Dynamics of objects	*Employee*	% of Employees present on May 2010 but not on April 2010 (new Employees) compared to the total number of Employees on May 2010.	3,0%
		% of Employees present on April 2010 but not on May 2010 (old Employees) compared to the total number of Employees on April 2010.	2,3%
	Employer	% of Employer present on May 2010 but not on April 2010 (new Employer) compared to the total number of Employers on May 2010.	2,5%
		% of Employer present on April 2010 but not on May 2010 (old Employer) compared to the total number of Employers on April 2010.	1,8%

It should be noted that the interpretation of these results is not directly connected to a quality evaluation, as a certain population dynamics is a characteristic of all phenomena. In the case of SSD, it is connected to demographic events for Employer unit and to hiring / termination of contract for Employee. However, the availability of the indicator values of each supply in time series together with possible reference value or benchmark could be very useful. For example deviations from an average or a trend value can detect the presence of possible errors in the supply analyzed.

4. Further development

The conceptual scheme just described and experimented on SSD should be implemented in an efficient manner as the evaluation of the statistical quality of the AD plays a crucial role in the new Istat statistical production process involving the use of AD. In order for the application to be effective, the standardized tools implementing quality indicators have to meet the following requirements:

- produce documentation of quality assessment in a timely manner;
- provide information as completely as possible;
- provide general information to AD users regardless of the domain of the produced statistics;
- be concise and easy to read.

In particular, there are three tasks that the QRCA can perform addressed to different types of users within NSIs.

The first one is “Evaluating AD statistical usability” for the new potential users of AD already acquired, or for AD sources acquired for the first time. For new potential users, the QRCA enclosed to the supplied AD, and together with the Metadata level description, could provide a useful support for evaluating whether to introduce or not AD into the production process. For new AD, after the preliminary metadata analysis performed using the Checklist proposed by Daas and Ossen (Daas and Ossen, 2011), the exploratory analysis provided by the QRCA at the Data level can give the information needed to support the final decision to include or not the source in the statistics production.

The second task of the QRCA is "Monitoring AD quality" already in use in NSI, for current users. This task is of primary importance because the production processes can develop a strong reliance on AD and a tool should be developed to promptly deal with possible problems. In particular it is necessary to constantly monitor AD quality 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 should be performed to make sure that there are no unexpected statistical errors.

The last QRCA task is "Monitoring quality in the AD acquisition process" that is to check whether data received are consistent with the requests and to support the process of data loading. Where appropriate, it is useful to define alert or warning to optimize the timing of data acquisition and release to internal users.

The quality evaluation results of the AD supplies in a time perspective can also provide interesting elements to evaluate the effectiveness of possible harmonization processes between administrative and statistical concepts agreed with the AD producers and the NSI such as: shared use of standardized classifications, changes in the process of recording data and so on.

The next challenge for Istat is how to plan and implement the quality reporting activity achieving the objectives defined and taking into account the limited resources available.
From the organizational point of view, Istat determined that the acquisition of AD should generally be made at a central level. A central organizational office, named ADA (AD acquisition and integration), is in charge of acquiring AD responding to almost all the institute AD requests. In 2013, Istat acquired about 250 supplies from more than 100 Administrative sources, so a strong coordination among departments using AD has been set up in order to plan the activities and meet the needs of the whole production process.

Recently, in order to avoid duplicate work among AD source users, this office also is building a new integrated system, called SIM (Integrated System of Microdata), which has the task to store AD supplies and to perform data pre-processing. In particular data received are coded with respect to official classification, when possible, and integrated using unique codes for the same objects in SIM. Currently unique codes are assigned to individuals and economic units. A Metadata repository is currently also under development. Of course, all operations are in compliance with the rules on data security and privacy.

The AD quality evaluation in SIM is the further task for ADA and this is another important function for Istat AD source users. From this point of view a standardized and generalized QRCA could be a support to share information defining usability of an administrative source and to monitor the quality of AD received by Istat (Di Bella and Ambroselli, 2014).

With the purpose of complying the appropriate timeliness, a system that allows to make the AD quality evaluation as automated as possible is being planning. Interesting results derived from the possible use of some statistical packages available in R (Tennekes et al. 2013). At this moment, in Istat, the implementation of the QRCA is undergoing testing on some education AD in SIM. The strategy aims to take advantage of all the available metadata, that is to make metadata "active" to the greatest extent possible for supporting the QRCA production process[14].

[14] Following the Core principles for metadata management (Common Metadata Framework Part A: Statistical Metadata in a Corporate Context), UNECE / Eurostat /OECD Group on Statistical Metadata (METIS) http://www.unece.org/stats/cmf/.

For the implementation of Source Hyperdimension quality indicators we'll try to take advantage of all the information used to manage the AD acquisition process.

In the Metadata Hyperdimension, we are experimenting some ways of interacting with the AD provider in order to acquire, together with data, also updated metadata necessary for their correct interpretation. In addition, the phase of Entity Relationship analysis of the administrative dataset and the consequent data loading in the relational database, can allow us to automatically identify the set of objects /entities to be evaluated.

The process of assigning an unique code to the same objects in SIM can provide information for the implementation of Comparability indicators of the objects in the Metadata Hyperdimension with respect to statistical units mapped in the system of data dissemination. In this case it could be possible to define equivalence classes for type of objects defined at different levels (i.e. individual-student, economic unit-enterprise).

In the Data Hyperdimension, a suitable description of the process of assigning unique codes can support the calculation of quality indicators for evaluating the record linkage procedure: some useful measurement methods can be derived for the linking variable quality indicator. It is important to underline that this is a core indicator not only for the Integrability Dimension evaluation but it also assumes a significant role for other quality indicators, such as Coverage and Dynamic of objects indicators.

A last example of making metadata “active”: the coding phase of the territorial units using the official classification in SIM, can produce Comparability indicators for the classification variable in the Data Hyperdimension.

5. Concluding remarks

The AD quality evaluation is a necessity for the statistical production processes and the QRCA is a useful summary, documentation and sharing tool.

The framework for describing the AD quality adopted has proved very robust in the different applications carried out (Daas et al., 2013) and it seems to be a comprehensive instrument including the many facets of the concept of AD quality with respect to their statistical usability. The ability to implement such a tool envisaging inter-operability of processes is interesting. From the first results of the implementation procedure, it follows that some indicators can be calculated automatically using the metadata process, while for other indicators, such as indicators of consistency checks (Accuracy of variables) it is necessary the source users contribution to define the check rules or, in case of first usability analysis, a collaboration with the team who is in charge of analysing the source for the first time.

The implementation activities are proceeding steps by steps and depending on the resources available, it will be possible to image a full or partial implementation of QRCA for AD in SIM. At the same time, “AD Istat user groups” are setting up for the most important data source holders (Tax administration, Social Security Institute, Ministry of Education, Universities and Research) in order to verify the possibility of sharing information or more specific analysis possibly useful to most users. In any case it will be important to spread the framework of the QRCA tool in order to standardize as much as possible the AD quality assessment process.

References

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

Burger J., Davies J., Lewis D., van Delden A., Daas P.J.H. and Frost J.M. (2011). *Guidance on the accuracy of mixed-source statistics*. Deliverable 2011/6.3 of Workpackage 6 of the ESSnet on Admin Data. http://essnet.admindata.eu/WikiEntity?objectId=5452

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://2011.isiproceedings.org/papers/950391.pdf

Cramer H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, USA, p282.

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

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. http://www.pietdaas.nl/beta/pubs/pubs/IJQR2011.pdf

Daas P.J.H., Ossen S., Tennekes M. (CBS, Netherlands), Zhang L. C, Hendriks C., Foldal Haugen K. (SSB, Norway), Bernardi A., Cerroni F. (ISTAT, Italy), Laitila T., Wallgren A., Wallgren B., (SCB, Sweden) (2011a). *List of quality groups and indicators identified for administrative data sources*, Deliverable 4.1 of Workpackage 4 of the BLUE-ETS project. http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.1.pdf

Daas P.J.H., Ossen S., Tennekes M., Zhang L. C. (CBS, Netherlands), Hendriks C., Foldal Haugen K. (SSB, Norway), Cerroni F., Di Bella G. (ISTAT, Italy), Laitila T., Wallgren A. and Wallgren B. (SCB, Sweden) (2011b). *Reports on methods preferred for the quality indicators of administrative data sources*, Deliverable 4.2 of Workpackage 4 of the BLUE-ETS project. http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf

Daas P.J.H., Tennekes M., Ossen S., Burger J., (CBS, Netherlands), Di Bella G., Galiè L., Bonardo D., Cerroni F., Talucci V., (ISTAT, Italy), Laitila T., Lennartsson D., Nilsson R., Wallgren A., Wallgren B. (SCB, Sweden), Hendriks C., Zhang L.C. and Foldal Haugen K. (SSB, Norway) (2013). *Guidelines on the use of the prototype of the computerized version of the QRCA, and Report on the overall evaluation results.* Deliverable 8.2 of Workpackage 8 of the BLUE-ETS project.
http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable8.2.pdf
http://essnet.admindata.eu/WikiEntity?objectId=5452

Di Bella G., Ambroselli S. (2014). *Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat*, paper presented at the European Conference on Quality in Official Statistics (Q2014) held Vienna, 3-5 June 2014.

ESSnet AdminData (2013a). *Final list of quality indicators and associated guidance.* Deliverable of Workpackage 6 of the ESSnet on Admin Data. http://essnet.admindata.eu/WikiEntity?objectId=5452

ESSnet AdminData (2013b). *Overview of Existing Practices in the Uses of Administrative Data for Producing Business Statistics in EU and EFTA* (Database Tables + Reference Library, update 31/12/2012). Deliverable 1.2 of Workpackage 1 of the ESSnet on Admin Data. http://cros-portal.eu/content/overview-existing-practices-uses-administrative-data-producing-business-statistics-eu-and

ESSnet AdminData (2013c). *Admin Data Glossary. Definitions adopted for certain terms related to the use of administrative data for producing business statistics*. Deliverable 2013/1.1 of the ESSnet on Admin Data. http://www.cros-portal.eu/sites/default/files//SGA%202011_Deliverable_1.1.pdf

Eurostat (2003a). Definition of quality in statistics. Working group Assessment of quality in statistics, Luxembourg, 2-3 October 2003. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/ess%20quality%20definition.pdf

Eurostat (2003b). *Standard Quality Report.* Methodological Documents, Working Group *Assessment of quality in statistics*, Luxembourg, 2-3 October 2003. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/STANDARD_QUALITY_REPORT_0.pdf

Eurostat (2005). *European Statistics Code of Practice for the National and Community Statistical Authorities* - revised edition 2011. Adopted by the Statistical Programme Committee on 28th September 2011. http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-32-11-955

Frost J.M., Green S., Pereira H., Rodrigues S., Chumbau A. and 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.

Hox J. J. and Boeije H. R. (2005). *Data Collection Primary vs. Secondary.* Encyclopedia of Social Measurement. http://joophox.net/publist/ESM_DCOL05.pdf

Laitila T., Wallgren A. and Wallgren B. (2011). *Quality Assessemnt of administrative Data.* Research and Development – Methodology reports from Statistics Sweden, 2, 2011.

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

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

Tennekes M., de Jonge E., and 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 Nation Publication, Geneva, 2007.

Wallgren A. and Wallgren B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons, Chichester, UK.

Zhang L.-C. (2012). *Topics of statistical theory for register-based statistics and data integration*. Statistica Neerlandica (2012), Vol. 66, nr. 1, pp 41-66.