Chapter 8: Quality

8.1 - Introduction – Users of statistical information mainly need to access easily objective information. While objectivity is not always an abstract concept, it is closely linked with both the users' needs (that is, to acquire information in relation to their own subjective needs) and with the purposes of who produces the information. Moreover, objectivity is also closely linked with transparency: an information is correct when it makes clear the aims, concepts, definitions, methods for data gathering and treatment, result aggregation criteria. Furthermore it is correct when it provides an indication on the quality of the aggregate disseminated. Transparency is what allows a user to easily comprehend the data, among those available, which can be used for his own needs. In this way the quality is defined as "*the degree to which a set of inherent characteristics fulfils requirements*"[1].

Referring to the Statistical Business Register (SBR) this definition means to consider its quality in terms of "*fitness of purpose*"[2] . The SBR purposes are to provide:

- information for the statistical units identification,
- the populations for sampling,
- the statistical outputs on the structure – in terms of units, economic classification, dimension - of an economic population and for business demography analysis,
- the tools for using administrative data for statistical purposes.

Users want Statistical Business Register (SBR) to be relevant, accurate and up-to-date. Relevant means that the SBR comprises all the units and the attached variables necessary to support the production of statistics. Accurate means that the information reordered corresponds to the reality. Up-to-date implies that the SBR provides the most accurate picture of the real world with the least possible time lag. Of course, the SBR quality is closely related to how it is used and to whether it satisfies the users' needs. However, different users resort to the SBR, and each of them have their own need. The SBR's reference universe and updating timing will be different if used for the *Short Term Surveys* rather than for the *Structural Business Survey*. For instance, if the *Value Added* is estimated based on the SBR's reference universe, clearly the quality of data (e.g. activity code and

[1] United Nations Statistics Division: National Quality Assurance Frameworks

[2] Jean Ritzen – Summary of quality session – 14th International Roundtable on BSF – Auckland, 2000

size) of large units will be fundamental due the weight that these units have in the estimation of such variable. On the other hand, if the indicators of the *Business Demography* take the SBR as reference, the quality of the smaller units will be very important due to their higher involvement in demographic events. Thus, if the SBR is a complex product because of its numerous and different users, the criterion for evaluating its quality will also be much more complex and it is not always possible to use the experience gained in measuring a standard statistical product such as a survey even more.

The consequence of survey managers requests impacts in conflicting demands regarding the timing SBR reports changes and is delivered. The solution is to adopt two versions of the frame, one being a frozen file for a certain period of time, e.g. one year, and one that reflects the latest available information. The frozen version(s) will serve for the time of sampling and at the time of coordinating the results; it is also used as reference population for business demography. The current file is continuously updated and more frequent version (monthly, quarterly) can be reproduced in order to report any correction of previous mistakes, changes in the more relevant variables (i.e. date of cessations, changes in addresses, changes in the relevant stratification variables) to support surveys according their timetable.

8.2 - *The quality components/dimensions in relation to the SBR* – The dimensions are "*the concepts used to describe some part or facet of the overall concept of quality, when applied to statistical outputs*"[3].

Using Eurostat criterion[4] the quality dimensions can be the following:

a) *relevance*, which refers to the degree to which statistics meet current and potential needs of the users.

b) *accuracy*, which refers to the closeness of estimates to the unknown true values;

c) *timeliness*, which refers to the period between the availability of the information and the event or phenomenon it describes;

d) *punctuality*, which refers to the time lag between the date of the release of the data and the target date (the date by which the data should have been delivered);

e) *accessibility and clarity*, which refer to the conditions and modalities by which users can obtain, use and interpret data;

f) *comparability*, which refers to the measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas, sartorial domains or over time;

g) *coherence*, which refers to the adequacy of the data to be reliably combined in different ways and for various uses.

A SBR must be pertinent to the needs of the users. In other words, it must contain *relevant* units and variables for constructing relevant populations and survey samples and/or lists for implementing these. It must be *accurate* in such a way as to reflect correctly the reality. As regards *timeliness and punctuality*, the timeliness with which the SBR is updated in order to reflect the events that occur in reality can be an important quality criterion even though it could be in conflict with the coherence

[3] United Nations Statistics Division: National Quality Assurance Frameworks

[4] Regulation of the European Parliament and of the Council on European Statistics, 14280/08 of 16 Oct. 2008, Article 12

and comparability criteria. However, the problem of updating the variables could be solved by keeping two versions of some variables, one version with the last variables updated and one with frozen values at a certain date (1 year). *Access* to the data could involve the possibility for internal users to obtain single data by directly connecting to the database, or for external users to obtain aggregate tables. Generally, the facility with which data of the SBR can be accessed must be considered as an important quality component. Another aspect of accessibility is the easiness with which the SBR information can be interpreted. Quality measurements, thus, involve the availability of documents necessary to correctly understand the information. As regards the *comparability***,** it is necessary to evaluate two aspects: *space and time*. In relation to space, since its comparability is ensured at a European level through a regulation, this component could be measured evaluating its level of adherence to such regulation. Comparability over time means to be able to compare both the units' data and the aggregates' data at different temporal periods. *Coherence* includes both *internal coherence* and *coherence with other registers***.** While internal coherence involves a consistent treatment of the register data, coherence with other registers is obtained using and archiving reference numbers. The use of a common identification code across all official business registers (administrative and statistical) is one way to obtain greater coherence. At last but not the least, even if it is not a quality criterion, cost is a quality constraint and can became a priority when allocating resources to improve any other quality aspect. Cost items concern burden on the data suppliers, often obliged to give many time the same information that they have already given to administrative bodies for other reasons: it follows as consequence a lowering in coverage, inaccurate answers, non response. Cost in terms of budget (questionnaires, interviewers, training, etc). Many of such costs can be reduced by extending the use of administrative data and in addition by adopting other coordinated tools to gather and update information such as a business portal.

***8.3 - Quality in survey and quality in the SBR: understanding the differences* –** The quality assurance is defined as: "all the planned and systematic activities implemented that can be demonstrated to provide confidence that the processes will fulfil the requirements for the statistical output"[5]. The identification of the specificities that characterize the statistical product "SBR" and their differences with the other "*standard*" statistical products - such as a survey - is a priority for the detection of a better SBR quality assurance and its components. Besides, the heterogeneity of the users of the SBR, its specificities and their impact on the quality issues, may be summarized as follow:

- ***Extensive use of Administrative data*** – The use of administrative data for statistical purposes has increased in the last decades. The SBR is the statistical product for which the "massive" use of administrative data is the main feature: all, or almost all, NSIs consider the administrative data as the priority sources to set up and up-to-date the SBR: thus the SBR quality is strictly connected with the quality of the administrative sources. Within a survey (or a system) for the collection of statistical data, quality is evaluated ex-ante and it is strongly linked to both the micro data collection process and the macro data production process. For example, accuracy of estimates is usually fixed in advance. On the contrary when using data stored in non-statistical (administrative) databases, for which statisticians do not have any control of the production process, the quality evaluation is set in a different context and it is resolvable only ex-post: *data are known but how they are generated is not*.
- ***Inputs Heterogeneity* –**The setting up and up-to-dating of the SBR need the use of more sources (administrative and statistical) to be integrated. Each source often provides partial information,

[5] United Nations Statistics Division: National Quality Assurance Frameworks

both with regard to the units and to the units' characteristics, so it is rarely sufficient to answer to all the SBR informative needs. Sub-population of units can be collected by different administrative sources (e.g.: very often the units in agricultural sector are stored in administrative sources different from the ones collecting the units involved in manufacturing and services activity); the characteristics of the units can be acquired from different sources (e.g.: turnover from VAT declarations, employment from Social Security Register); information on the large and complex units are collected using different statistical techniques (surveys or profiling), while information on the smallest units are collected using – almost always – only administrative data. In such a way the classical theoretical model used, that connected one source (the survey) to one informative need, is not suitable any more. The model should be transformed in "an informative need – different sources", that implies the identification of new and specific methodologies for the treatment and the quality evaluation of the information coming from a variety of sources. Global quality evaluation of the SBR is difficult to obtain. It would be better to split the SBR information in different pieces and to develop a set of quality indicators able to evaluate each subset (each phase of the process, each partition in terms of units or variables, etc.) and afterwards trying to develop a complex (composite/compound) indicator, in order to evaluate the global quality. Furthermore the heterogeneity of the inputs implies the development of specific criteria to evaluate the *internal coherence* of the SBR.

- ***Inputs Variability over time*** – In the case of a statistical survey, the stability of contents and process is almost always guaranteed, or it is anyhow easily to be kept under control. Most of the problems that arise when using administrative sources is connected to the changes adopted in the source itself that are not known: changes in the classification criteria, in the registration and cancellation rules, in the used administrative control processes. Substantially, "big" changes occurring during a certain period (year) must be considered not plausible. Therefore, the main objective is to verify the "stability" of the sources and to avoid that merely administrative changes produce "non-real" structural changes in the SBR.
- ***Relevance of technological aspects* -** The process for setting up and up-to-dating the SBR is characterized by:
  - use of a huge amount of data,
  - development of complex procedures for data integration and methodologies implementation,
  - changes over time in applied rules due to changes in classification, in admin sources contents, adding new information, etc.

  The industrialisation of the process is a relevant element of the SBR. The control of the technological elements in the SBR process and the evaluation of the quality of the technology used (software and hardware) is a central task in the global quality assurance of the SBR.
- ***Output specification*** – The main objective of the SBR is the dissemination of individual data to be used by the surveys as sampling population or target population. The dissemination of micro data suggests that "errors annul each other on average" is not true anymore. With reference to SBR in terms of quality evaluation, the errors add one to another one: e.g. there is not a "coverage" error but an "over- *plus* under-coverage" error to be evaluated.
- ***Users heterogeneity*** – The quality policy adopted by the SBR must take into account users' need. As sampling frame for supporting STS and SBS surveys, the BR update processes have to be as timely as possible and the produced frames have to guarantee stable reference universe. At the same time SBR has to ensure the most complete coverage possible. In addition, accuracy of certain variables (e.g. activity code and size) of large units will be fundamental for the estimation of surveyed statistical variables (i.e. Value Added) while the quality of the smaller units will be very important when calculating indicators of Business Demography. Another important aspect

of quality is coherence. Having all business surveys using the same conceptual and physical frame leads to a more coherent and cohesive business statistics. It is much easier to reconcile the statistics coming from the various business surveys programs for the System of National Accounts when all these surveys start from the same base. Whereas for internal users quality should be guaranteed at micro level (for example a wrong address of a peculiar unit may also have legal implication) for other users outside NSIs it is relevant to assure consistency in time series, accuracy at sectorial or regional level in some aggregates directly derived from the SBR.

8.4 –The peculiarity of SBR: the continuous updating of the SBR data - Updating a unit in a statistical business register implies how to identify and treat the actual changes in the variables occurring over an established time lag. Such changes can involve the unit existence, unit characters and the ties between the recorded units. These events have to be referenced at a specific time period. According to the basic accounting equation, population of units in the SBR at time t+1 is given by:

$$N_{t+1,k} = N_{t,k} + B_{\Delta t,k}$$

where $B_{\Delta\ t,k} = I_{\Delta\ t,k} - O_{\Delta\ t,k}$ determines the incoming (I) and outgoing (O) flows of units over the period (*t, t+1*) where k be the total or only a sub-set (by size, sector, geographic area). Incoming/outgoing flows are determined by i) birth/death of enterprises, ii) changes in the classification characteristics.

The updating of the register at time *(t+1)* is not only determined by actual changes, but also by adjustment of characteristics of the units. SBR data cannot be considered like the results of a statistical survey, because the latter is a process that couldn’t get any additional information once it is concluded, whilst the former, during the period (*t, t+1*), may acquire data referring to a previous time, even earlier than *(t)*. Therefore in the SBR it is possible to modify information referring to the time *(t)*, like the NACE classification of the unit, the measure of its size or the date in which a cessation or a birth is registered, particularly when this event is caused by merger or demerger among firms. In this way it is possible to adjust errors – wrong classification, wrong size of a unit, etc. – or assessing the correct referring period in which an event has happened. Sometimes actual changes are recorded at a later time, though they apply to an earlier period, and there could be a delay in recording birth/death or in recording changes in characteristics in the administrative registers used for updating the SBR. A dynamic and productive updating of a SBR to a large extent depends on the way in which the administrative archives register the information that they receive from the enterprises. There are often errors and delays especially when delivering data for small enterprises.

Taking into account the different kind of adjustments (actual and error changes), the continuous updating of SBR determines the possibility to revise any set or sub-set of information produced with regards to the state of activity (determining the false active units or on the contrary the false non-active ones) and the main classification variables like the economic activity code or the number of employees. A missed detection and certification of adjustments between *(t)* and *(t+1),* acted on data referring to *(t)*, could lead to spurious modifications, that cause in their turn false reading and understanding of the evolution of the economic system. The identification of the two components of error i.e. the balance of the errors (or adjustments, or delays in updating) related to the state of activity of units and the balance of the errors related to the other characters that delineate the scope of a target population if a fundamental task for a better evaluation of the SBR quality.

8.5 - The role of quality of the administrative data sources for the SBR

NSIs are increasingly making use of administrative data for statistical purposes. As the major input of information of SBR are administrative data, the Business register quality has to be evaluated taking into account the entire process of acquisition, loading and processing of administrative records. This extensive use of administrative data is predominantly stimulated by the need of reducing response burden and costs, and at the same time, to increase the coverage of the frame or subset of sub populations, and the completeness of some variables. The increase in the use of administrative data sources despite of the numerous advantages determines its dependency on external data sources and relates with the quality of those sources. It is common to use different strategies of updating BR data, and data from the administrative archives are essential to update the archive of small-medium enterprises, for which it is impossible to obtain direct statistical information since any inability to sustain such costs. The administrative sources that feed the BR, their frequency and the amounts of records are generally the first information that the BR team intercepts to have a feeling of how quality is.

To manage quality in advance, in the input step of the process, when selecting the data sources to assign values for the main variables in the SBR, the decision is predominantly based on the quality of the sources containing information about those variables. Sometimes statisticians are obliged to use the only available source (i.e. VAT turnover from the VAT declaration taken from fiscal registers), other times many values exist to choose the correct one among them (different addresses available in different sources), some other times the administrative variables have to be used in an indirect way as a proxy of the needed statistical variable. According to the general framework (ref: Eurostat: QUALITY ASSESSMENT OF ADMINISTRATIVE DATA
FOR STATISTICAL PURPOSES, 2003)

Among the basic statistical requirements that administrative data must fulfill, the following are relevant:

1. availability of metadata about the administrative dataset contents. Metadata must describe the administrative procedures that create the data, any important administrative events relevant to the data and definitions of concepts, variables and the population they refer to;
2. administrative data should be relevant for the units to be covered and/or for the variables suitable for the purpose. Units and variables must be coherent with the statistical ones;
3. administrative data files must contain identification variables that allow the link and integration with other sources. The presence of the same unique key in all files should be the best mean to facilitate record matching.

A way to compare and analyse the administrative sources is to standardize the determination of their various quality components (The quality framework developed for registers, Daas et al., 2009). An overview of the quality components are referred to as hyper dimensions (Karr et al., 2006), and are called: Source, Metadata, and Data. As framework for the quality analysis each hyper dimension is composed of several dimensions of quality and each dimension contains a number of quality indicators.

Source and Metadata quality assessment is usually done by doing evaluation as scores given to different key dimension describing the Sources used and related metadata.

The hyper dimension of Source considers all those activities that allows to exploit the quality of the information contained in each of the data source than one want to use on a regular basis. The most common quality dimension concern the Frequency of delivery (yearly, monthly, continuously), the relevance with respect to the needed information and its satisfaction of demands, the relationship with the supplier and the procedures which evaluation reveal the dependency risk from data provider.

The Metadata hyper dimension focuses on the conceptual and process related quality components of the metadata of the source. Prior to use, it is essential that a statistical office fully understands the metadata related quality components because any misunderstanding highly affects the quality of the output based on the data in the source. Metadata deals with the clarity of changes into legal environment that affect the administrative definition and changes into forms of acquisition of information, comparability of variables when the time period variables cannot be transformed easily into the required statistical variables time (i.e. weekly variables or average to be transformed into time points). Comparability of values because differences in the reporting periods of sources. Another relevant quality evaluation concerns identification keys, it considerably hinders combining this data source with the other sources of information.

Finally, the evaluation of the quality of the data, the Data hyper dimension focus on quality indicators that can be developed to describe, in a quantitative or qualitative manner, the quality of input in the statistical process. They refer to Technical checks, Accuracy, Completeness, Time-related dimension and Integrability.

8.6 - Frame errors and their implication on surveys – The Eurostat manual on Business Register define an error as "*a difference in the information presented in the register and the information as it should be, according to a chosen image of the real world produced and maintained by an accepted instrument and documented procedures*[6]". The manual define the following type of errors:

1. *Errors in existence* - This type of error is due to false information regarding the demographic variables (date of creation and date of cessation) for a particular unit. There are two categories of existence errors:

- Units are recorded as economically active, but are not yet, or no longer, active in the real world. This results in over-coverage, and can lead to response problems for statistical surveys based on the register.
- Units are economically active but are not present in the register. This type of error results in under-coverage, and can also adversely affect the quality of register outputs.

2. *Errors in identification variables* - Errors in names, addresses, telephone numbers etc. can hamper data collection due to problems locating and contacting the units. Errors in names and addresses also impede the use of statistical business registers as tools to link and co-ordinate data from different sources. Errors in legal form are similar in some respects to the errors in stratification variables considered in the next paragraph. They can affect the inclusion of units in certain register outputs, and in certain strata of survey samples.

3. *Errors in stratification variables* - This type of error includes errors in variables such as the economic activity code, size-class (number of persons employed, turnover or net assets) or the geographic area in which the unit is situated. These errors result in inefficient sampling and strata allocation for surveys based on the register, and will be, to some extent, detrimental to population estimates derived from the register.

[6] EUROSTAT – Business Register Recommendations Manual – Chapter 18 – The treatment of errors.

The errors that occur in SBR have e big impact on the process and results of the survey based on the register. It is known that the purpose of each survey is to produce estimate as accurate as possible of a given unknown parameter. Sampling and non sampling errors determine the level of quality of sample-based estimates in fact they cause bias and a loss of efficiency. Among non sampling errors non responses and coverage problems in the frame of reference represent the main sources of error.

These two factors are correlated because some non-responses can be attributed to errors in the frame such as the impossibility to contact the unit included into the target population as well as an incorrect information in the frame determines the necessity to delete some unit in the sample reducing its size. The evaluation of the impact of the frame errors on the estimates of a survey is a direct measure of the accuracy of a SBR

Frame errors and their impact of the overall error can be classified according to the following types:

a) under-coverage – SBR does not reflect businesses within scope for that survey. Reasons for

under-coverage errors are enough well known: omission (lags and leakage), errors in the determination of the state of activity of units (falsely not active units), mistakes in stratification variables (out of scope units when they are in scope). SBR under-coverage generally affects estimations increasing bias.

b) over-coverage - SBR considers in scope businesses that are not. Reasons for over-coverage are the opposite of the under-coverage ones: duplication, errors in the determination of the state of activity of units (falsely active units), mistakes in stratification variables (in of scope units when they are not in scope). Over-coverage generally affects estimations increasing their bias; moreover if a sampled unit is correctly identified as ceased, a reduction of the sample size determines an increase in the sampling error.

A specific attention has to be given to errors due to an incorrect information held by units correctly registered. Coding errors typically affects stratification variables such as principal economic activity codes, size in terms of employment, location variables, demographic data. This type of errors produce inefficient sampling and strata allocation.

When a unit is located in a place different from that registered in SBR the results is an increase of the total non-response rate (in particular for postal surveys). The impact of this error is both on bias (a respondent unit will represent the missing one but it can significantly be different) and sampling variance (reduction of the sample size).

8.7– Metadata

The utilisation of metadata are requested in every phase of any statistical business process. Metadata consists of properties directly derived from data and documentation to better comprehend how data was generated. The metadata system is formed by some items such as descriptions and definitions of statistical data and variables, used classifications, variable formulas and unit of measurements, document and product metadata about suppliers, publication information, identification knowledge of the publications or products, field or subject area glossary, keywords, technical information of data and variables which are used in producing process. Metadata help users to comprehend about the BR quality.

In the SBR metadata can take different forms, a used schema is the following:

- The source of the data: in the SBR units and variables derive from the integration process of statistical and administrative sources. Moreover they derive from the activity of continuous updating by clerical staff, using online information on internet or directly contacting the businesses and thus correcting errors on the base of the results of an editing and imputation procedure. Source codes generally consist of alphanumeric codes assigned to each unit and variable to indicate the source of information. They can be used to identify exactly which source gives the information on the identification characteristics, for example whether a specific variable like turnover is taken from the VAT declaration rather than from the SBS surveys, whether the employees is directly estimated from the social security files or it is taken from the survey on large business units.

- the procedure used for character attribution (imputation model, estimation, directly from survey, etc.);
- the production process (survey, integration of administrative registers, direct contact);
- whether changes occurring in the period are variations or adjustments;
- reliability of data (with reference to the generating process and sources);
- The date of updating or history of each data: , these can be linked to data items to indicate the date to which the data relates and/or the date on which that particular data item was last updated in the register. The more recent the more reliable;
- Any documentation about sources and processes is vital in helping users to assess the quality of register data. It can added to disseminated information or take part of the database (web application) to be consulted by any user.

8.8– The tools for quality evaluation – It is possible to identify tools and/or actions able to assess the quality of the SBR. Some of these tools are "generic" because applicable for all statistical products, some others are specific for SBR.

- *The users' survey.* If the quality is connected to the users' needs, the knowledge of the perception on the SBR that the users have must be the first tool to be used for the SBR quality assessment. The objective of an users' survey (US) is to collect information on satisfaction of the users of the SBR. The aims of the US will be to collect information on quality dimensions: relevance, timeless and punctuality, accessibility and clarity with reference to the "main" users of the SBR corresponding to the "business surveys" that use the register as sample or target frame. Very often the evaluation of these quality components could be different from one statistical domain to another one (e.g. Structural and Short terms statistics can identify different "information relevance"), moreover the 100% of satisfaction for all of the users is not realistic. The results of the US have to be weighted using certain evaluation of the user's "relevance" within the whole national statistical system.
- *Auditing.* "*A quality audit is a systematic, independent and documented process for obtaining quality audit evidence (records, statements of fact or other information, which are relevant to the quality audit criteria and verifiable) and evaluating it objectively to determine the extent to which the quality audit criteria (set of policies, procedures or requirements) are fulfilled*"[7]. Auditing is a "powerful tool…..by providing important information"[8] to improve the quality of the SBR. Since the use of the external audits could be expensive for the NSI, it could be desirable to conduct an internal audit using a team of auditors – not in charge of the SBR process – that both users of SBR and statisticians without any (or only a partial) knowledge of the statistical registers contest should attend.

[7] United Nations Statistics Division: National Quality Assurance Frameworks

[8] Eurostat – Data quality assessment and tools

- *Register quality survey.* The realisation of a SBR quality survey is an useful tool – even if expensive – to acquire indicators on register accuracy especially of the variables “location”, “economic activity code” and “size”. To avoid self referential findings, the survey results must be analysed in an independent way by officers different from the SBR ones. Furthermore the results must be handled carefully because of the difficulty to find out a “real” value to compare the SBR contents.
- *Auditing of clerical work:* Quality audits are a useful tool for monitoring the quality of clerical processing and automatic updates. They can be achieved through regular analyses of key variables or clerical checks of a representative sample of update actions or, preferably, a combination of the two approaches. According to an analytical approach Quality audits consist in monitoring changes in the register and can be done on a regularly base or at least during the period just before the dissemination of the BR for users. As a starting point, frequency distributions for some key variables can be compared before and after an update (automatic or manual) to assess the impact of the changes on different subsets of units and to ensure that all changes can be adequately explained. Generally it can be organized by priorities giving a higher weight to large relevant units or to relevant units for specific sectors or under a particular observation from users. Another way for audits is take samples from the list of clerical updates to monitor the quality of the clerical input to a register. These Sample checks should be regular, as random as possible in nature, and should also cover a representative selection of updates. Clerical audits are normally undertaken by experienced staff, who investigate the work of BR staff to see if the actions taken comply with the current guidelines. Or, it is done by cross checking the clerical results, assigning the same units to be updated to different clerks and then comparing their results. The rate of clerical errors can then be monitored over time and reduced, for example by improving the training or by addressing the staff when possible. Both approaches should be closely linked, preferably producing regular summary reports to inform managers and users. In order to be fully effective, the quality audit function should be closely linked to the documentation and training functions in a form of ‘quality circle’ so that issues identified are resolved, documented and covered in future staff training. This should make it possible to ensure sustained improvement in quality over time.

Process and data checks – An important tool is the editing and imputation process. n order to check for errors or inconsistencies, micro editing and imputation procedures have to be developed. A quality check plan must take into consideration: the main relevant variables (i.e. the state of activity of the unit, the legal status, the economic activity code, the number of employees), their links or relation, the development of edit rules and the action to take to impute or correct data. In any editing procedure the hierarchical order of variables to be checked is always taken into account. Some typologies of errors can be solved easily with automatic treatment by deterministic rules. Such rules can avoid to commit logical errors in uploading data i.e. formal validity dates of characters or other recurrent errors.

Other rules ascertain for the same unit (at micro level) any inconsistency among different variables at t or the same variable having two different values between t and t-1. Edits are often classified as fatal edits (or hard edits) which determine the need of a correction and query edits (or soft edits). In the first case correction can be automatic, after having identified the best rules, or made by expert staff. Query edits tend to have the overall control of group of units i.e. the high size ones or specific sectors when a change has been detected between a time period to another. For future usage and in order to document, all checks and corrections feed the database according by recording the history of checks. In this DB all metadata information about the process of editing, the origin of data, the corrected value the reference dates, the kind of imputation are recorded and determine a reduction of punctual operations in following years.

8.9-The conceptual framework for quality indicators

The implementation of a system of quality indicators in BR is a very common tool to asses quality. Before calculating any quality indicator (QI) it must be clear the ***key factors for defining a BR quality indicator:***

- Time: The BR is not a fixed object in time then a QI is characterised by a reference date (t) or a reference period ($_{t-\ t+n).}$In presence of a lag between the reference period (t) and the examination period the indicator will describe the BR quality at *t* measured at *t'*.
- Scope: QI is applicable to a given set of units: i.e. by type of units (enterprises, local units,…) and form a subset in the overall set of possible units. It is formally defined by a filter presented in the form of a logic formula operating with the register's variables (e.g. active enterprises born before t-x)
- Subpopulation: Inside the units under scope, QI must be applied to sub-populations of interest; it is not useful to create "a global" q. indicator: it may mask weaknessess in a specific sub-population. It is useful to define indicators, at *least*, by - size (small/medium/big), territorial areas, - sectors of activity
- Variable: QI must be applied to a given variable of the register
- Criterion: To construct a q. indicator it is necessary to have a criterion for estimating, unit by unit, the quality of the variable. For each unit and for each variable it must be possible to assess whether the value is correct: right/wrong (true/false), a scale or degree of quality between 0 and 1

There are different **criteria** to assess quality:

a) External information sources - A value of a unit in the SBR can be considered as correct if it is sufficiently "close" to a reference value (external sources). It is the most commonly used criterion, that focuses on *compliance* (whether a value of a variable in the the SBR complies with the value of the same variable in an external source) as a proxy for *reliability* (whether the value of a variable is exact or not) as it is the only possible measure when the "real" value is not known. The *compliance rate* (% of units for which the variable assumes the same value) replaces the *reliability rate* (% of units for which the variable is exact)

b) Register survey (proving survey) - The same is true as for the previous scenario: the individual quality is assessed by comparison with a reference value. This practice is expensive.

c) Internal consistency: A value will be deemed "correct" if it is coherent with the other variables of the same unit (turnover/employees, main activity/legal status). Definition of consistency edits is difficult and often they are just "plausibily" edits .If the variable passes the edit there is no guarantee that it is correct

d) Temporal consistency: the quality of a variable can be defined on the basis of comparison with its previous values in time series. The aim is to define impossible or less plausible changes from one period to another.

e) Quality without "witness": it is possible to identify a set of information able to access quality without needing a reference value and with no element of comparison:

- the information validity date: the date on which the information was most recently checked or updated "more recent the information is, the better it is" ;
- the name of the information source;
- the methodology adopted;
- other metadata identified by the BR staff.

8.10– The SBR quality indicators in the overall BR process –Examples

The quality of the register can be ensured only when verifying the quality of the register's input (sources of origin) and the quality of the processes used for treating and integrating such inputs. When implementing quality assurance activities where the output (the SBR) is the result (process) of a set of integrated admin or statistical sources (input) three aspects must be taken into account:

1. Quality of the INPUT;
2. Quality of the process (matching, merging, editing, updating);
3. Quality of the OUTPUT

Quality of the Input

SBR is a highly heterogeneous product because of many inputs sources. Then the quality of the administrative sources affects the SBR quality and becomes a very useful indicator itself. Nevertheless, it is impossible to control ex-ante the quality of each administrative archive from the SBR's point of view. Rather, the quality of the source can be evaluated only ex-post by means of suitable analyses to identify any error in the supply of data and to adopt adequate corrections through integration processes. The first approach toward an evaluation of quality involves the accessibility and clarity of administrative data, that is the ease with which the BR updating process can access data. For instance when any change in the format(s) in which data are available are precisely reported or when there exists enough metadata, illustrations and accompanying advices. Then strong efforts have to be done to make the management of metadata transparent.

Thus, simple indicators of the quality of the source are:

➢ A – Time lag: difference between the date on which the data are supplied and their reference year.

➢ B – Indicators of completeness of the variables: for each variable (company name, address, etc.), it is possible to calculate the ratio: 1-Nmissing/Ntotal

Most of the problems that arise when using administrative sources regard changes adopted in the source itself that are not known: changes in the classification criteria, in the registration and cancellation rules, in the administrative control processes used. Substantially, "big" changes that occur during a certain period (year) are considered as impossible or as not very plausible. Therefore, the main objective is to verify the "stability" of the sources and avoid that merely administrative changes could produce "non-real" structural changes in the BR.

Simple indicators are based on the comparison of the values provided in two different years:

➢ C – weight % of the variations (per character); once a synthetic indicator (median) of the number of variations has been selected, further analytical verifications are carried out when the weight of the variations in a year exceeds the average level.

The fact of comparing the subsequent supplies of data from the same source is also fundamental for analysing the completeness of the enterprises' creation and cessation dates. Such analysis is essential to avoid problems of under-coverage or over-coverage of the SBR. In particular, when a simple indicator, that counts overall the cessation dates of enterprises during a reference year, does not provide particularly relevant information, then only an analysis on time consistency will allow evaluating its quality. A complex indicator is then used to evaluate the informative loss or gain and to identify the data concerning the enterprises' cessation dates:

- D – The loss of information for update delay of the source is obtained by comparing, if available, the values of a variable - for example, the number of cessation in year t – in both the yearly supplies received in year t and in the following year t+1 – referred to the year t = 1-Ncess(t+1)/Ncess(t). In this way the lag in the registration of the cessation dates in the input source could be estimated. As regards the cessations, clearly, if a shorter time for the acquisition of administrative data is chosen (the t-th supply=reference time), an important share of information can be missed. This indicator should be compared with the same value calculated for the succeeding year t+2, to see if the gain of information is not increasing in a significant way. Thus, it seems inefficient to wait for a longer period (t+2) in the hope of some improvements in the registration of the dates.

Quality of the Process

A complex system like the one producing an SBR, requires the process be maintained under control through the use of some quality indicators calculated ad hoc for that phase of the process.

Any administrative sources based SBR production and updating process presents some critical points such as the complex system of logical and physical integration of records,–the procedures for estimating units' characters, the editing and imputation plan. The process can be logically divided into three macro-phases. Each of them can be described with associated indicators controlling the quality.

Macro-phase 1: the integration of input records from administrative sources

The purpose of the first macro-phase of the SBR production process is the integration of administrative archives and the creation of clusters referring to the same entity (the enterprise). On the inside, two different sub phases can be distinguished and then described by quality indicators.

First sub-phase: Link intra-archive – Inside each input source, records that pertain to the same legal entity are integrated (where the legal entity is represented by the Fiscal Code). Synthetic quality indicators are:

- A – weight % of fiscal codes duplicates (by source) / total number of supplied records, (temporal consistency). A decrease of this indicator over time indicates an increase of quality;
- B – number of new records (with respect to the previous t-1 supply) by source in the year t. It indicates a coverage measure with regards to unit creations. In particular, by comparing the weight % of new records using time series, a graph is useful to compare trends at source level and among different sources. Moreover, by comparing this indicator with a benchmark (for example with official statistics for the enterprise demography), sudden variations in trend for ratios could mean potential quality indicators of the source.

Second sub-phase: Link inter-archives – Integration of records coming from different sources and related to the same unit in order to build-up a "cluster of records" for the same enterprise. Again, group of records represent the same legal unit if they pertain to the same fiscal code. The administrative benchmark that is the base used to integrate all the other sources if the Fiscal source. This phase is very peculiar having as aim of identifying the set of administrative information available for each legal unit. Mistakes that could be done in this phase – missing or wrong links – can affect considerably the results coming from the following steps. Quality indicators are:

- ➢ C – Number of clusters of records in year (t) with the presence of the Fiscal register (reference for the BR year t) (1)
- ➢ D - Number of clusters of records in year (t) without records from Fiscal register (2)
- ➢ E – Number of clusters of records in year (t-1) without records from Fiscal register (3)
- ➢ F – Under-coverage indicator : [(3) ∩ (1)]/(1)

Macro-phase 2: The estimation of characters

The aim of the second macro-phase of the BR production process consists in the imputation of the main attributes to each unit and in the identification of active units in year t. Each attribute's estimation procedure can be evaluated using ad hoc quality indicators. These indicators make use of outputs produced step-by-step by the implementation of the procedure.

In particular the setting up of the frame of active units is of high priority in the process: other characters are checked only for active units that define the reference universe for sampling and for the economic structure. Compliance rates can be calculated with respect to the main sources that hold information strictly related to the active/not active status of unit.

- ➢ G – Percentage of concordance/discordance rates calculated between the SBR status by source indicating active/not active units (i.e. SME surveys, structural changes database), active units (i.e. Foreign trades survey) and not-active units (i.e. bankruptcy database)

Macro-phase 3: editing and imputation procedures

The third macro-phase of the BR production process relates to the editing and imputation process and aims to obtain the final identification of the universe of active units for the reference period year t. In order to measure the quality of the adopted edits and procedures, quality indicators can be calculated. Some of them measure the amount of errors produced by each rule.

The check plan is usually developed as a group of projects working as separate modules to be executed in order; they can be also changed in number and composition. Each project is characterized by a set of common rules having a similar structure and affecting data in the same direction.

The main projects are:

- Cleaning – rules determining the exclusion of some units from further checks
- Deterministic – if/then clauses, that cause the automatic change in the values of the involved characters, whenever the conditions occur
- Errors - assessment and errors rules, that cause the warning for a follow-up whenever the conditions occur.

Some remarks need to be done. Since the number of edits could be very high, usually deterministic rules focus on peculiar subset of information and on specific characters (for example, the cross combination of Nace code and size). Editing process allows to produce warnings of supposed errors. Also in this case there is the need to contain the number of warnings reducing the number of edits that can be checked. Trained staff concentrates controls only on meaningful larger units instead of less economic significant ones.

The calculation of quality indicators based on the number of warnings or errors, edited and imputed by each project, is misleading. Appropriate indicators can be built looking at trends over time that measure the increase/reduction of units involved by each type of edit. A synthesis can be done as follows:

- ➢ H – Variation (%) between t and t+1 of units involved by type of error

- I - Variation (%) between t and t+1 of units automatically changed by deterministic rules
- L - Variation (%) between t and t+1 of units having a warning and manually verified

8.11 – Improving quality

The following are examples of ways in which the quality of the SBR can be improved. These examples or suggestions concern one or more different quality component at the same time, since it is clear that they cannot be improved separately as there are trade-off that need to be balanced each time.

First of all a survey frame assessment tool for the preparation and co-ordination of survey and for grossing up survey results must be developed and maintained In order to allow each survey to monitor their survey populations, this tool should provide a directory from which mailing lists can be assembled for the despatch of questionnaires in statistical surveys; it must provide a population of businesses for which efficient sampling schemes can be designed and panels monitored, since every survey area needs to understand the changes happening to their population of interest between survey cycles then tracking births, deaths, arrival, departures, and other significant changes; it must provide the basis for grossing-up results from sample surveys to produce business population estimates.

Improving Timeliness

In order to have the most representative image of the business population possible, the BR updating process have to be as timely as possible, assuring at the same time the more complete coverage. Regular frame updates can be done by systematically applying updates which are available from the relevant administrative data that in the national system are responsible for inscription and dilation of businesses (Fiscal for the opening of VAT for operating businesses and Chamber of commerce) important to catch creation, cessation and events of structural change. As well as using the number of employees from Social Security monthly data. Coverage by Administrative information such as sales tax remittances, income tax returns, and payroll deductions provide clear signals when a business is active. Signals of when a business becomes inactive are much less clear. Another aspect is to reduce the time required to apply updates so that these changes can be reflected in the survey universe as soon as possible. Often surveys' managers comply that a change they receive from questionnaire are not immediately adopted to update the business register population but time is spent from BR staff to ascertain the real change and perform updates. A way could be to give grant permissions to subject matter specialists to perform the updates, after having received the appropriate training. Another possibility is to let the supplier (the business) to update directly their info by using a business portal.

Improving completeness

It is desirable to extend the BR to more information about the businesses characteristics required by surveys, to hook their surveys to the business register.It should be possible to link the BR to satellite registers, like an employment database or some specific sectors satellite database to be used to directly update the secondary economic activity within the BR, thus maintaining its accuracy, relevance and overall quality level.

Improving coverage

Different strategies must be set up for under coverage (with respect to a certain economic activity sector is due to misclassification), and for over coverage (usually duplication of the same units). The risk of duplication depends firstly from the usage of input sources and matching procedures. In presence of common codes sometimes the risk is to misinterpret the code, i.e a fiscal code is sometimes (rarely) confused with the Vat code, as in most of the juridical entities they are the same. In absence of a common code, the use of Record linkage techniques is suggested. To this regard many software have been developed and have high performance, however each automatic procedure must be adapted to the Country's peculiarity; for example in the registration of enterprise names or in the managing of addresses, the treatment of such identification characters used in matching must be taken into account.

Improving quality reports

Reporting is crucial to disseminate about quality. The aim is to deliver intra-annual quality reports (i.e. monthly) to alert users of the main changes over time. Changes can affect the top level universe, employment, turnover and number of businesses. Drills down into industry, region, legal status, administrative sources. Annual quality reports can track of small changes each month and can show trends comparable by year. It is then possible to monitoring peculiar cases such as: foreign controlled transport companies (employment can be underestimated according to the company policy) having discrepancies between employment/turnove, big enterprise group in construction field operating abroad, Holding companies.