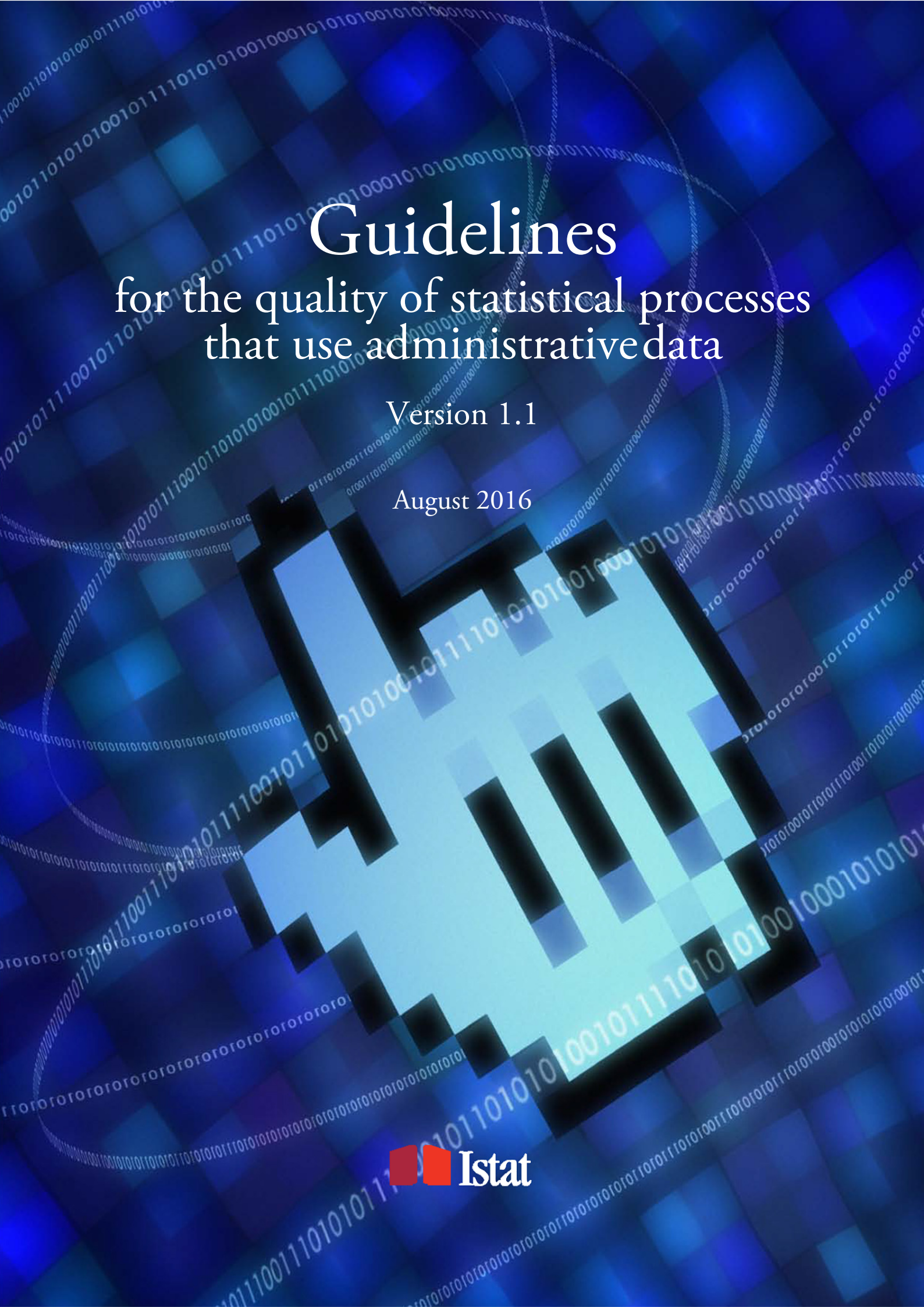
Guidelines

for the quality of statistical processes that use administrative data

Version 1.1

August 2016



Istat

The draft of this manual was coordinated by G. Brancato

The following people contributed to the draft of these Guidelines:
Brancato G., Boggia A., Barbalace F., Cerroni F., Cozzi S., Di Bella G., Di Zio M., Filipponi D., Luzi O., Righi P., Scanu M.

This manual was revised by:
Signore M., Di Bella G., Di Consiglio L., Filipponi D., Luzi O., Pallara A., Rapiti F., Simeoni G.

We would like to thank the Quality Committee for their collaboration.

TABLE OF CONTENTS

Introduction

The increasing use of administrative data for statistical purposes creates the need to adapt the quality assessment standards and tools in order to take account of its peculiarities. This has oriented towards the draft of these ***Guidelines for the quality of statistical processes that use administrative data***, which aim to integrate with the items concerning the use of administrative data, those already produced for survey-type statistical processes and already published on the ISTAT website in Italian and English[1].

The main goal is to equip the Institute with a reference for the conduct of processes using data from administrative sources and an instrument for their evaluation with a view to audit or self-assessment.

These guidelines are structured into three parts concerning input quality, the quality of the process that uses data from administrative sources and output (or product) quality, respectively. They follow a conceptual model described in Section 1. The Part A regards the principles and guidelines on the acquisition of data that are input of the statistical production process and the measurement of their quality.

In the Part B on the quality of a typical statistical production process that uses data from administrative sources, for each phase of the process itself, a principle representing the aim to pursue is stated and the methodological guidelines for its fulfilment provided.

The Part C, product quality, concerns the quality of output resulting from a production process that has used administrative data, and focuses attention only on products of "macrodata" nature. As with the previous guidelines, this part is not organized into principles, but proposes again the Eurostat definitions of quality, and describes how the use of administrative data can affect their meaning. Finally, definitions of the errors, which occur in the different stages of the production processes that use data from administrative sources, are given.

The increased use of administrative registers for statistical purposes has led many National Statistical Institutes, including the ISTAT, to the identification of permanent or temporary structures (departments, committees and commissions), strategic and operational, with different functions and various levels of decision-making, in charge of supporting the managerial and methodological activities associated with the use of such data. These functions range from: raising the awareness of owners of administrative data, to coordinating and monitoring variations in the administrative forms, to the internal needs review, to the management of the relations with the owners and acquisition of registers, to the documentation and evaluation of the quality, to monitoring on internal uses. The principles and guidelines on the centralised acquisition of administrative registers are developed in the attached Appendix.

These guidelines are addressed to the managers of the statistical production processes, who have to fully understand the aspects related to the quality of a process and of the statistical product that uses data from administrative source, and who should also be aware of the problems that the acquisition of administrative register poses for a National Statistical Institute.

These guidelines have some overlapping elements with those developed for statistics surveys, elements that were considered useful to have in order to complete the volume.

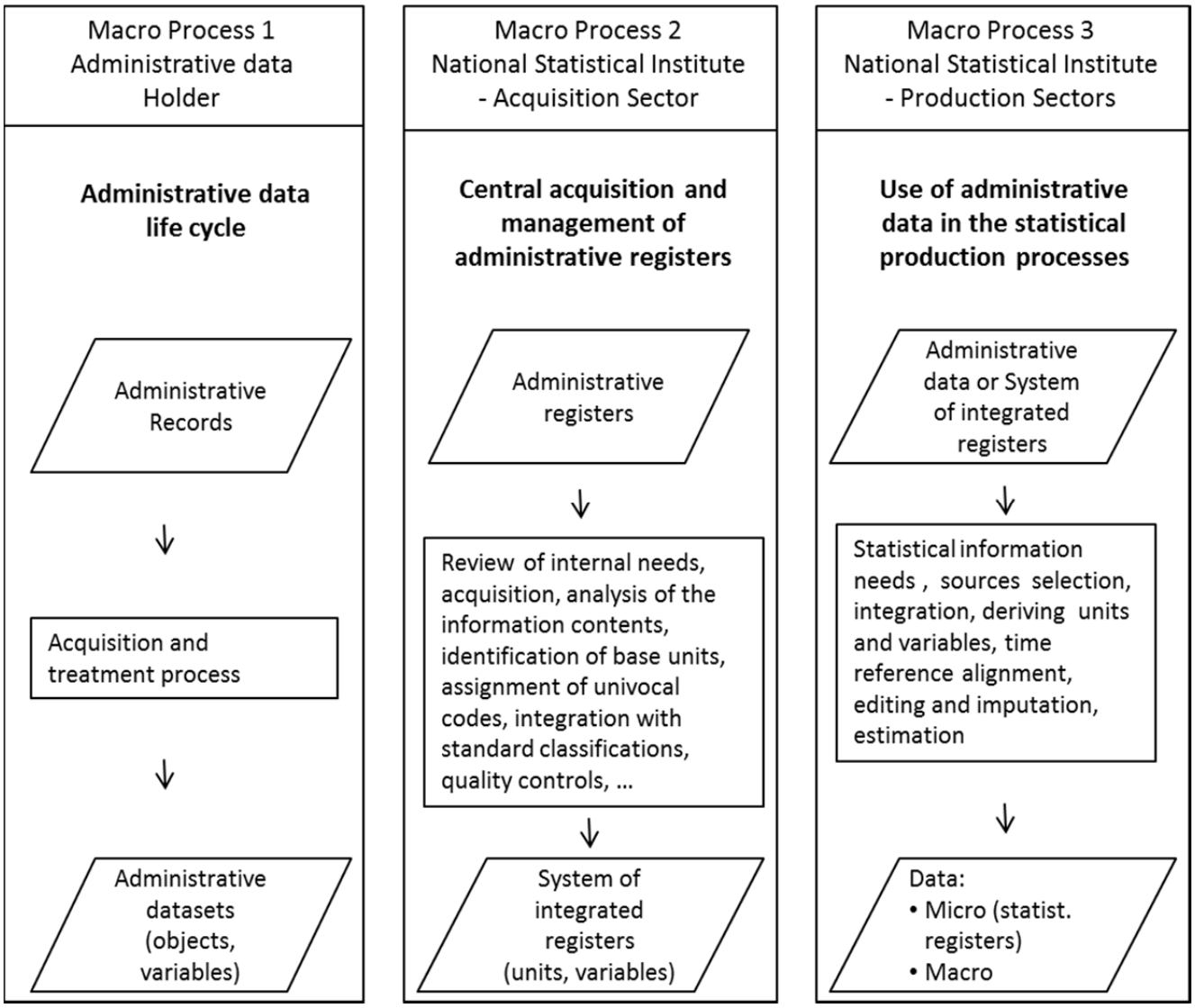
[1] English version available at the following link http://www.istat.it/en/files/2011/11/QualityGuidelines_EngVers_1.11.pdf

1. Framework for the quality of statistical processes using administrative data

The use of administrative source data in the statistical production processes introduces specificities that require a revision in the representation of the production processes as well as in the errors that are generated, compared to the traditional direct survey-type production processes.

The model underlying these guidelines identifies three main macro processes regarding administrative data, briefly shown in Figure 1, where for each macro process the in charge subject, the information input, the transformation process and output are shown.

Figure 1. Macro processes on administrative data



The macro process 1, of responsibility of the administrative data body, can be likened to a survey data collection and, as such, its errors can be modelled according to the classic survey approach (Groves *et al*., 2004; Zhang L.C., 2012). This macro process is generally beyond the control of the National Statistical Institutes and therefore is not specifically covered by these guidelines. However, it should be noted that the same institutes may have a role of coordination and harmonisation of the forms, as it happens in our country,

aimed to increase the acquisition of administrative data and their usability for the production of statistics. For this reason, more important than the quality of the data, in this setting reference is often made to their usability for statistical purposes.

The macro process 2 reflects a trend seen in an international context (see for example Wallgren & Wallgren, 2014 and Statistics Finland, 2004), which considers the acquisition of administrative archives as a process with a strong level of centralization, oriented to the construction of an integrated system (or a set of integrated systems and/or systems that can be integrated) that responds to internal needs. This is configured as a real process of acquisition, treatment and release of an output, which represents an intermediate product of a statistical nature, and in turn an input for other statistical production processes (macro process 3 in Figure 1). The administrative data acquired at this stage are subjected to a set of procedures mainly aimed at singling out the statistical objects starting from the administrative objects and supplying them with the necessary information for their subsequent use. At this stage, aspects of a managerial nature concerning the relationships with the administrative data holders and issues related to the metadata documentation supporting the use of the acquired data, assume the utmost importance. The quality indicators included in the hyper-dimensions "Source" and "Metadata" of the checklist developed by Statistics Netherlands (Daas *et al*. 2009) are an example of measurements for the quality monitoring and assessment. Principles and guidelines concerning this macro process are dealt with in the Appendix.

Finally, the macro process 3 concerns the statistical production processes that use data from administrative sources that have already been centrally acquired and pre-treated (thus available from the macro process 2) or administrative data obtained directly from the administrative data holders (thus available from the macro process 1). The processes belonging to the third macro process are finalised to the production of macro-type statistics (estimation production) or micro-type data (statistical register production). For each of these processes, the underlying quality model identifies:

i) input quality (administrative data and/or system of integrated registers in Figure 1)
ii) the quality of the process using administrative data (through-put quality)
iii) output quality (micro and macro-data)

Input quality
Since the input comes from a collection process that is exogenous from the Institute control, it is of utmost importance to check the quality of the administrative data source acquired, taking into account the specific statistical production objective (see Part A). Input quality dimensions and indicators, oriented to the output, have been developed within the international project Blue – ETS (Daas P., Ossen S., 2011).

Through-put quality
With respect to process quality, phases and relative sub-processes, information objects and potential errors on units and variables are shown in Figure 2. It uses the current standards for metadata, such as the *Generic Statistic Business Process Model* - GSBPM (UNECE, 2013a) and *Generic Statistical Information Model* - GSIM (UNECE, 2013b), customizing them for the use of administrative data. For the identification of errors, some elements were taken from the Zhang work (2012). In GSIM informational objects are identified in generic form, and are characterized according to the stage of the process in which they are "outputs." For reasons of clarity, in the diagram generic objects are specified by the type they take on. With regard to errors it is important to note that sometimes there may be propagation both in sub-processes that are different from those in which they are generated, and between the various entities (units, variables).

Figure 2. Main phases, sub-processes, informational objects and potential errors for the variables and units

Main phases and sub-processes (GSBPM)	Information objects - GSIM (unit)	Potential errors (unit)		Information objects - GSIM (variables)	Potential errors (variables)
1. Specify needs 1.1. Identify needs 1.4. Identify concepts 1.5. Check data availability 2. Design	Population (statistical target) ↓ Observable administrative data or system of integrated registers	No conceptual correspondence (under-coverage, over-coverage)	Coverage error	Target concepts ↓ Variables (represented and/or derived)	Specification errors No conceptual stability
4.Collect * 4.2. Set-up collection 4.3. Run collection 4.4. Finalise collection	Observable administrative data or system of integrated registers ↓ Acquired administrative data or system of integrated registers	Selection errors (missing units, duplications, delays)		Variables (represented and/or derived) ↓ Instance variables **	Measurement error
5. Process 5.1.Integrate data 5.5. Derive new variables & units 5.9. Align time references*** 5.4. Edit & Impute	Acquired administrative data or system of integrated registers ↓ Population (statistical analysis)	Linkage errors Unit derivation error		Instance variables ↓ Data set (validated microdata)	Processing error (misclassification, missing data, intra-source and inter-source consistency errors, implicit or explicit model assumption error)
5.7. Calculate aggregates				Data set (validated microdata) ↓ Estimates and/or macrodata	Implicit or explicit model assumption error

* Can be an acquisition from the body in charge of the administrative data or it can be an internal transmission from a centralised unit in charge of acquisition and pre-treatment'

** GSIM terminology *instance variable* has been adopted.

*** This sub-process is not present in GSBPM.

One of the first tasks to be tackled in using administrative data for specific statistical purposes is a careful analysis of the correspondence between administrative concepts and the statistical objectives of measurement, together with the planning of the aspects related to the acquisition of the archive or subset of interest data (sub-processes 1 and 2 in Figure 2). Data collection is differently characterised if a transmission procedure, internal to the Institute, is assumed or if a direct acquisition from the body in charge of the administrative data is considered. The conceptual non-correspondence with the statistical objective (in terms of population and the object of observation) limits the validity and usability for statistical purposes of the administrative data. The columns on the units show that the search for population of statistical interest could lead to the acquisition of populations derivable from several administrative archives.

Moving on to the columns on variables in Figure 2, when the data from administrative source do not correspond to the concepts underlying the target variables, it is referred to as specification error, likewise for the surveys with questionnaire (Biemer and Lyberg, 2003). Moreover, any changes of legislation or procedures governing the administrative act over time or across space limit the time or the geographical comparability of the concepts and therefore the comparability of the data time series.

If administrative data are used to partially or completely replace the units to be collected by direct surveys, it can be assumed to be a phase similar to that of collection/acquisition, during which the concepts underlying the administrative register become data observed and interpreted in terms of statistical variables. As well as in the direct surveys, observation may generate measurement errors; this same type of error may exist when using administrative data for statistical purposes (column relative to the variables in Figure 2). As regards the units, the data acquisition is associated to a potential error, which is given the name of “selection error”. Nonetheless, this is an error that actually affects the coverage; in fact, any problems in the data acquisition and any delays in the register updates, can lead to the selection error and consequently to coverage errors.

The majority of types of use (construction of statistical registers, integration of survey data concerning subpopulations or sets of variables) foresees, inside the process "5. Process", an integration sub-process, that can be achieved through record linkage or matching procedures. In this phase errors such as false matches and false non-matches can be generated at unit level. The first can in turn generate under-coverage in the units and inter-source consistency errors in the variables, e.g., due to variables pertaining to a unit being imputed wrongly to different other units. False non-matches, however, can result in over-coverage and, if the integration is aimed at filling values in the variables, can also lead to missing data on the variables.

In the activity of derivation of new units we talk about possible derivation error that may concern the same unit or its relations with other units. Derivation errors can cause coverage errors. The activities on the derivation of new variables and time references alignment (sub-processes 5.5. and 5.9.) regard simple coding or complex reconstructions made on the variables and can cause errors which were called "classification errors." Special attention is required to minimise the potential errors which can be generated during the necessary transformations to align the data time reference to the intended one.

The editing and imputation process aims to improve the data quality and to solve problems of internal consistency to the sources and, in general, does not constitute a significant source of error. However, the generation, in this phase of a potential editing and imputation error, is conceivable, due to an erroneous specification of the rules, the use of a wrong model or to errors in the correction activities carried out manually.

At the end of the treatment phase the population of interest is reconstructed and, if compared to the target one, it would allow the final coverage assessment. However, it is worthy noticing that the target population is rarely available.

Nonetheless, it should be stressed that with the current practice it is not always easy to distinguish the individual error sources attributable to the various sub-processes of integration, derivation of the units and variables and editing and imputation, because they often involve procedures which are conducted simultaneously.

Finally, during the estimation process, the most significant errors can be those from model assumption because the statistical production processes using administrative data commonly resort to the model-base theory.

The principles and guidelines to limit and control the errors generating during the statistical production process are stated in the Part B of these guidelines.

Output quality
As regards the quality of the macro-type output (or product quality) the approach followed is based on the quality dimensions defined at European level (Eurostat, 2009) and adopted by Istat (Part C).

Concerning the micro-type output, the model here developed envisages the quality evaluation during the statistical production process, specifically after the validation phase, and when the validated microdata file is stored, issue tackled in the last principles of Part B.

Some bibliographic references

Biemer P.P., Lyberg L. (2003). Introduction to survey quality. Wiley, New York.

Groves R.M., Fowler F.J.Jr, Couper M., Lepkowsky J.M., Singer E., Tourangeau R. (2004). Survey Methodology. Wiley, New York.

Statistics Finland (2004). Use of Register and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland.

UNECE (2013a). The Generic Statistical Business Process Model GSBPM v5.0 http://www1.unece.org/stat/platform/display/metis/The+ Generic+Statistical+Business+Process+Model

UNECE (2013b). Generic Statistical Information Model GSIM v1.1. http://www1.unece.org/stat/platform/display/metis/Generic+ Statistical+Information+Model

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. Second edition: John Wiley & Sons, Chichester, UK.

Zhang L.C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistics Neerlandica, Vol. 66, No. 1, pp. 41-63.

Part A. Input data quality

Principle A.1. Acquisition of administrative data

The data from administrative source necessary to the statistical production process should be obtained directly from the Institute units in charge of their centralised acquisition, when they are already available from these units. If the administrative data to be used are not in the acquisition plan, then their acquisition from the owner bodies should follow, as much as possible, the Institute standard procedures.

Guidelines

During the planning phase, the statistical production units declare their needs of administrative data to the unit in charge of the centralised acquisition. The latter arranges the collection and analysis of these requirements, setting priorities if the case. Indeed, it could not be possible to centrally acquire all the requested administrative registers, and it could be necessary to concentrate resources on a subset of registers that have strategical relevance.

If case the administrative data necessary to the statistical production process are already available in the Institute, they should be obtained exclusively through internal transmission. To this aim, the relative procedure, generally stated in internal protocols/agreements, should rigorously be followed, applying the appropriate rules on security and privacy. It is additionally important to provide feedback to the units in charge of the centralised acquisition on possible quality deficiencies of the transmitted data.

If the administrative data necessary to the statistical production process are not part of the centralised acquisition planning, and it is allowed their acquisition directly from the owner body, it is advisable to follow as much as possible the standard acquisition procedures adopted by the Institute. In detail, the following elements, among those widely defined for the centralised acquisition (see Principle 3 in the Appendix), assume particular relevance:

- the identification of people within the administrative body for the data transmission;
- the establishment of formalised agreements setting the data transmission timing, the expected quality levels of the register, the documentation supporting the transmission of the register;
- the compliance to the rules and procedures concerning the protection of confidentiality which ensure the transmission methods and the treatment of sensitive data and prevention against the risk of confidentiality breach.

More in general, good relationships with the body providing the administrative register should be established and maintained as well as collaborations with the aim of continuously improving data quality through-out the exchange of information with the body itself.

Some bibliographic references

Statistics Canada (2009). *Statistics Canada Quality Guidelines*, Fifth Edition – October 2009

Statistics Canada (2009). *Statistics Canada, Use of administrative data (website)* http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm (ultimo accesso: Dicembre 2013)

United Nations Economic Commission for Europe - Unece (2011). *Using administrative and secondary source for official statistics: a handbook of principles and practice*. United Nations, New York and Geneva, 2011.

Principle A.2. Quality assessment of input data

Input data quality, whether obtained by a centralised unit or directly from the body owner of the administrative data, should be measured and assessed with respect to the specific statistical objective, before their treatment and integration in the statistical production process.

Guidelines

The first quality control, with respect to the specific predetermined statistical objective, should be applied to the input data, whether they have been acquired from the owner body or transmitted by the internal unit.

The element to be considered are various and only for some of them it will be possible to compute quantitative measures.

Some issues that can impact input data quality pertain to the administrative background, e.g. with respect to the regulations and the stability of the procedures ruling the production (legislation, forms, …) over time and geographically. The procedures of recording of the administrative events by the owner body and the timing for the furniture of the register to Istat, as well as the quality of the documentation provided to support the administrative data, may also have impact on quality.

Coverage of the administrative register or of registers (or a subset of data drawn from registers centrally acquired) should be measured with respect to the specific statistical target populations underlying each single register.

With regards to the variables contained in the registers or in the datasets drawn from registers centrally acquired, in the first place, a conceptual analysis on the correspondence with the statistical target variables should be carried out, in order to prevent specification errors. Afterwards, it is opportune to compute data quality measurements, e.g. the extent of missing items in the variables of interest.

In case of data acquired by the centralised unit in charge of the acquisition, some indicators of interest for the specific statistical objective could already be computed and made available.

Some indicators on the quality of the administrative data used as input, in a view oriented to the output, are available in Daas e Ossen S. (2011).

Some bibliographic references

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

Daas P., Ossen S. (2011). *Report on methods preferred for the quality indicators of administrative data sources*, Blue – ETS Project, Deliverable 4.2.

Part B. Process or through-put quality

Principle B.1. Information needs and choice of administrative sources

The information needs to be met should be clearly defined. Each decision on the use of administrative data, according to the stated statistical objectives, should be preceded by an overall assessment of the characteristics and quality of the data contained in the source. In the case of availability of multiple sources, the choice should be made on the basis of a comparative analysis.

Guidelines

The identification of information needs requires, as is commonly the case for direct surveys, in-depth knowledge of the users and of their information needs, knowledge to be acquired through the establishment of user-producer continuous and stable dialogue. In general, users are heterogeneous and often have conflicting interests: it is thus important not only to know the various user types, but also to be able to rank user relevance with respect to process results. As a consequence, the main users should be clearly identified, involved in the object definition and (re)planning of the process and their satisfaction should be measured, with various levels of formalisation and involvement. It is useful to produce and regularly update documentation on main users and their characteristics.

The information objectives and the uses of administrative data (direct production of statistics, construction of statistical registers, quality support to survey-type processes), expected within the statistical production process, should be identified and planned in advance in order to identify the quality requirements of the administrative data and the most suitable treatment methods.

In the case of direct production of statistics, through the replacement of units and/or survey variables with administrative data the target population and the statistical units, the variables and the classifications of interest, the geographical and time dimensions assume particularly importance. With respect to these elements, it is advisable to carry out an objective analysis, first predominantly conceptual then more oriented to the data, on the usability of the administrative data in the statistical production process, for the specific purposes.
Firstly, starting from the definition of the target statistical population, namely from the specification of the units that compose it, the time references and the geographical boundaries (e.g., the population resident in Italy at a certain date), a careful analysis should be made of the ability of available administrative datasets to fully reflect the units of the population. Likewise, the correspondence between concepts and definitions that concern the statistical variables of interest and those regarding the variables inferred from administrative data should be evaluated. In this regard, the quality aspects on which to focus attention concern the coverage levels expected for the population of interest, the validity of the variables used and, therefore, the evaluation of possible bias arising from inappropriate use of the administrative data. These aspects will be further investigated in the following principles B3 and B4.

In the case of construction of statistical registers, when several administrative sources are used together in order to ensure the greatest possible coverage of the population of interest, particular importance is assumed by the quality dimension relative to data integrability, that should be assessed in the first place with respect to the existence and quality of matching keys and/or univocal identifiers, and then in relation to errors which may be generated in the *record linkage* procedures (see Principle B2 on Integration).

When data from administrative sources are also used to support the statistical production processes, such as: sampling frames, information source to improve the efficiency of the sampling design and the estimation phase, assist in the editing and imputation phase, assessment of quality and of specific sources of error (e.g.

in the process of validation of the results or for estimating coverage errors), their quality has a significant impact on the quality of data produced and, what is more important, depending on the specific use, are the aspects of timeliness of the administrative data, the accuracy of the information, its consistency and comparability, the integrability of the datasets.

In general, for statistics and/or statistical registers to be produced with regularity and continuity, particular relevance is assumed by the aspects concerning the frequency, timeliness and stability of the administrative source data used.

Where there is availability of several usable administrative data sources, unless it is decided not to acquire them all to integrate them, the decision of the best source is to be made on the basis of a comparative analysis of the advantages and disadvantages of using one source rather than another and on the expected impact in terms of quality of the produced data. In this sense it may be useful to conduct scenario analyses, and especially to carefully consider the risks associated with situations of unavailability of the source, identifying alternative strategies to ensure statistical production.

The general assessment on the quality of administrative data sources to be included in the production process and the comparative analysis between the sources should be based on all the indicators of the input quality (see Principle A2 and Daas P. and Ossen S., 2011).

The assumptions and motivations that led to the choice of the administrative source data to be used and the type of use within the statistical production process should be properly documented.

Some bibliographic references

Cerroni F., Di Bella G., Galiè L. (2014). *Evaluating administrative data quality as input of the statistical production process.*Rivista di Statistica Ufficiale, issue no. 1-2, 2014

Daas P., Ossen S. (2011). Report on methods preferred for the quality indicators of administrative data sources, Blue – ETS Project, Deliverable 4.2.

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

Lavallée, P. (2000). "Combining Survey and Administrative Data: Discussion Paper." ICES-II, Proceedings of the Second International Conference on Establishment Surveys, Survey Methods for Businesses, Farms, and Institutions. Buffalo, New York. June 17-21, 2000. pp. 841-844.

Statistics Canada (2009). Statistics Canada Quality Guidelines, Fifth edition (Chapter on the Use of Administrative Data) http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.htm

Principle B.2. Data integration methods

Administrative data should be linked once the objectives of the linkage activity have been clearly established, by means of sound and commonly recognized methods. The linkage process should be clearly described, pinpointing every non-testable assumption. All the phases should be tested and documented. The validity of the results of the data integration process should be evaluated and documented by appropriate indicators.

Guidelines

When talking about the use of administrative data in a statistical production process, data integration is a sub-process whose aim can be, roughly speaking, twofold:

- On the one hand, the objective can be the development of an integrated microdata dataset that can be further used for macrodata production purposes by increasing the number of variables that can be jointly analyzed;
- On the other hand, the objective can be population completion, e.g. for developing a reference list.

Data integration can make use of different procedures. If the data sources to link share an identifying key that can be considered as error-free, it is possible to use exact matching methods. When this identifying key is not available in one of the data sources to integrate, it is possible to select a set of statistical variables that jointly allow to identify the units of the population of interest (e.g. name, surname, address, …). In this last context record linkage procedures are used: they are based on the comparison between these variables for the different units observed in the two data sources to integrate. Possible errors in these variables increase the possibility to generate errors in the linked data set, affecting the results on the use of administrative data.

The record linkage procedures and the necessary steps for their application do not take into consideration the nature of the data sources to link, whether two administrative sources, or an administrative source and a survey. The record linkage steps will be shown in the sequel with the necessary reference bibliography (a detailed, though not complete, description of all the steps is in Scanu, 2003 or Herzog *et al*., 2007; for a description of the same problems under an information technology point of view, see Batini and Scannapieco, 2006). All the record linkage steps as well as the underlying hypotheses should be clearly documented. Furthermore, the data integration process should be performed fulfilling the norms preventing any disclosure risks.

Choosing a record linkage approach: the deterministic and the probabilistic approaches. Who is performing a record linkage has two distinct alternatives: choosing a deterministic or a probabilistic method. As a matter of fact, a final linked data set can be obtained by successively applying different record linkage procedures, including the possibility to start with a deterministic approach and then add other linked pairs of records by means of a probabilistic approach. Anyway, the probabilistic and deterministic approaches are substantially different.

Deterministic approach: the mechanism under a deterministic record linkage consists of a set of rules that declare in a clear way which pairs of records are declared as matches and which pairs are not, by a simple comparison between the key variables. An example of a deterministic record linkage, frequently applied in practice, consists in declaring as matches those pairs of records having the same personal identification number (ID code, fiscal code, social security code,…). If a record linkage procedure consists only of this rule, the underlying assumption is that this key variable is not affected by any kind of errors.

Probabilistic approach: while in the deterministic approach the rules, and hence the matches, are chosen by who is conducting the record linkage, in a probabilistic record linkage the objective is to estimate the probability that two records can be declared as matches or non matches, having as a clue the comparison between the key variables. This implies that the results of the comparisons between the key variables are observations generated by two distinct probabilistic models: one for the matches and the other for the non-

matches. These two probabilistic models are rather different, i.e. the intersection between the sets of values these two models generate with probability larger than zero will be narrower the better is the quality of the matching variables. As a limit case, when the matching variables are not affected by errors, the model characterizing the matched pairs will assume with probability one the equality of the matching variables, while the non-matches assume the same value with probability zero. If the matching variables are affected by errors, the sets of values that the two models are able to generate have a non-empty intersection. This non empty intersection is the cause of the record linkage errors: false matches and false non-matches. The probabilistic record linkage procedures have the objective to determine a set of record pairs to be declared as matches by assuming under control (i.e. defining in advance) the probability to commit the record linkage errors (false matches and false non-matches) in such a way that those pairs whose status as a match or non-match is undeclared (and that should be submitted to costly procedures) are the lowest possible. The models and the procedure have been defined by Fellegi and Sunter (1969). The decision that a record pair is a match or a non-match is based on the likelihood ratio test, i.e. the ratio between the likelihood that the observed record pair is a match, given the observed comparison between the matching variables, and the same likelihood for non-matches. The larger this ratio, also called "weight", the more likely is that the observed pair is a match; on the contrary, low values of the ratio suggest that the hypothesis that the observed pair is a non-match is more likely. Intermediate values of the ratio characterize those pairs where it is difficult to take a decision, because the hypotheses "the pair is a match" and "the pair is a non-match" are almost equally likely, and they should be analysed by well-trained clerks (clerical review). In general, the decision that each pair is declared as a match or a non-match is taken by fixing two thresholds, an upper and a lower threshold; pairs whose weight is larger than the upper threshold are declared as matches, pairs whose weight is under the lower threshold are declared as non-matches, pairs whose weight is in-between the upper and lower thresholds should be submitted to clerical review, an expensive (in terms of time and money) activity. Hence, the declaration of a pair of records as a match or non-match is associated to the tolerance limit that have been fixed by means of the probabilities to commit mistakes, i.e. the probability to declare as a match a pair that is a non-match (false match) and the probability to declare as a non-match a pair that is a match (false non-match).

Choosing the matching variables. Among the common variables in the two data sources to link, the selection of the matching variables should primarily aim at considering those variables that are jointly able to identify the single units of the population of interest. A suggestion is to include those variables that are not affected by errors, missing values, lack of stability in different reference times, confidentiality problems (National Statistics 2004a, 2004b). For the record linkage applications related to households and persons, Gill (2001) describes some suggestions. Statistics New Zealand (2006) suggests to avoid as matching variables pairs of highly correlated variables. The final objective of the matching variables is to discriminate as much as possible pairs of records that are matches against those that are non-matches. In case of categorical variables, this is likely to happen when the number of categories is large: special cases are those where some categories are highly discriminant because of their rarity, see Winkler (1989, 1993, 1995, 2000).

Choosing blocking and sorting methods. Record linkage procedures compare each pair as available in two data sets: hence, the number of comparisons is equal to the data set sizes product. This number can be too large for both computational and statistical reasons (Scanu, 2003). In this framework, before performing a record linkage procedure it is advisable to partition the files according to one or more high quality categorical variables, comparisons are limited to those pairs of records that belong to the same partition (Baxter *et al.,* 2003). The partitioning variables are named *blocking variables*, and their high quality avoids the loss of potential matches that are not detected by the procedures because of a disagreement in a blocking variable. If high quality variables are not available it is possible to use the *sorted neighborhood* method by ordering the variables according to one or more variables (e.g. in alphabetic order) and comparing pairs of records that are similarly ranked in the two orderings (see Batini and Scannapieco, 2006).

Choosing a comparison function. The matching variables are chosen among the common variable in the files to link, after having chosen the blocking and sorting variables. These variables are chosen for their predictive power on the status of a pair as a match or a non-match, i.e. they are such that the pairs of records with slight differences tend to be considered as matches, while larger amount of differences make more likely the hypothesis that the pairs are non-matches. The detection of the amount of difference between the matching variables in a pair is computed by a comparison function. There are many comparison functions, whose objective is to pinpoint some results when comparing the matching variables, see Gu *et al.* (2003). Anyway, the most used comparison function is the one that computes the equality (1) or difference (0) of the values of each matching variable as observed in the records to compare.

Modifying the record linkage results under the 1:1 constraint. Both a deterministic or probabilistic approach declare the status of each single pair as match or non-match, disregarding what happens in the other pairs. This approach can give contradictory results if some constraints hold: for instance, the procedure can be constrained so that each record in a data set can be linked to no more than one record in the other data set. As a matter of fact, a record linkage procedure as described up to now can declare as matches two pairs that share one record in common. In order to solve this problem it is possible to use the transportation algorithm (Schrijver, 2003) whose objective is to minimize the sum of the weights of the pairs that are definitively declared as matches under the constraint that a record in a data set can be matched to a maximum of one record in the other data set.

Quality evaluation. Any (deterministic, probabilistic, mixed) data integration procedure should always be evaluated through the computation of the two main indicators on the possible record linkage errors: false matches and false non-matches. Often there are not any kind of tools for performing this kind of evaluation, in other words more accurate sources where it is possible to verify the correctness of the declared matches. This means that it is necessary to manually analyse the results, at least on a sample of pairs. Another aspect that influences the validity of the procedure is the assumption of non-testable hypotheses: e.g. the rules applied in the deterministic approach or the hypotheses underlying the likelihood function in the probabilistic approach are not always suitable for the data at hand (for instance it is not true that a comparison variable on a pair is generated independently from and equally distributed to the other pairs). See Belin and Rubin (1995) for a discussion on these aspects. In the last years Tancredi and Liseo (2011) have defined a Bayesian record linkage method that tends to solve some of the problems underlying the application of the Fellegi and Sunter approach.

Some bibliographic references

C. Batini, Scannapieco, M. (2006). Data Quality: Concepts, Methodologies and Techniques. Springer Verlag, Heidelberg.

Baxter R., Christen P. and Churches T, " A comparison of fast blocking methods for record linkage", Proceedings of 9thACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation, 2003

Belin T.R, Rubin D.B. (1995). A Method for Calibrating False-Match Rates in Record Linkage. Journal of the American Statistical Association, Vol. 90, pp. 694-707.

Fellegi, I. P., and A. B. Sunter(1969). A theory for record linkage. Journal of the American Statistical Association, Volume 64, pp. 1183-1210.

Gill L. (2001).Methods for automatic record matching and linkage and their use in national statistics, National Statistics Methodological Series No. 25, London (HMSO)

Gu L., Baxter R., Vickers D., and Rainsford C. (2003). Record linkage: Current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, Australia, April 2003.

Herzog, T.N.Scheuren, F.J. and Winkler, W.E.(2007). Data Quality and Record Linkage Techniques. Springer Science+Business Media, New York.

National Statistics (2004a) National Statistics Code of Practice - Protocol on Data Matching. Office for National Statistics, London.

National Statistics (2004b) National Statistics Code of Practice - Protocol on Statistical Integration. Office for National Statistics, London.

Scanu, M. (2003). Statistical methods for record linkage. ISTAT, Collana Metodi e Norme, n. 13. [ISTAT, Collana Methods and Standards, no. 13.]

Schrijver, Alexander (2003). Combinatorial Optimization - Polyhedra and Efficiency.Springer Verlag

Statistics New Zealand (2006). Data integration manual; Statistics New Zealand publication, Wellington, August 2006.

Tancredi A., Liseo B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. Ann. Appl. Stat, Volume 5, Number 2B (2011), 1127-1698

Thibaudeaus, Y. (1993). The discrimination power of dependency structures in record linkage. Survey Methodology, Volume 19, pp. 31-38.

Winkler WE. (1989). Frequency-based matching in the Fellegi-Sunter model of record linkage, Proceedings of the Section on Survey Research Methods, American Statistical Association, 778-783.

Winkler, W.E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. Proceedings of the Survey Research Methods Section, American Statistical Association, 274-279.

Winkler, W.E. (1995). Matching and record linkage. Business Survey Methods, Cox, Binder, Chinappa, Christianson, Colledge, Kott (eds.). John Wiley & Sons, New York.

Winkler W. E. (2000). Frequency based matching in the Fellegi-Sunter model of record linkage (long version).Statistical Research Report Series no. RR 2000/06, U.S. Bureau of the Census. At http://www.census.gov/srd/papers/pdf/rr2000-06.pdf (last view, 21/04/2008).

Principle B.3. Units identification and derivation and coverage assessment

The procedure for identifying and deriving the statistical units should follow established practices. All assumptions should be explicit and the steps should be documented. The quality in terms of coverage should be properly assessed.

Guidelines

The identification and reconstruction of the statistical unit is an issue rarely mentioned in the classical literature concerning survey techniques. This is justified by the fact that in the survey design, the population, the statistical unit and the reporting unit find correspondence in planned observation. This may no longer be valid when the target population is observed from administrative type sources. In this case, the statistical units of interest do not always exist as such in administrative datasets and the administrative population does not always coincide with the statistical one.

Therefore, particularly when (re)planning the process, it is necessary to study the objects contained in the administrative dataset and their relationships with the units that are considerable for statistical purposes and the assessment of the representativeness of the statistical population by the administrative one.

Identification and derivation of the units. The objects of an administrative dataset can be events or administrative units and their relationship with the statistical units is not always immediately identified.
The statistical units are created through derivation[2] by administrative objects (Wallgren & Wallgren, 2014) using a transformation function, which makes it possible to align the administrative data with the statistical one, first on the level of metadata (through comparison and reconciliation of definitions) and then at the data level by making explicit the treatment to be applied the administrative data in order to use them for statistical purposes. Unfortunately, given the heterogeneity of administrative datasets, often their treatment does not follow standardised methods and procedures, but is specific to the statistical objective and the acquired source.
It is important in the derivation process of the new units to carefully evaluate the applicability of the techniques available in the literature (Wallgren A. and Wallgren B., 2014).
In order of increasing complexity, the reconstruction of the statistical unit starting with the administrative unit can be: (1) simple, (2) assisted by expert, (3) assisted by integration with other datasets, or (4) mixed.

The statistical unit reconstruction is said simple when the administrative unit coincides or is easily attributable to the statistical one by aggregation of administrative units based on defined criteria. For example: (a) individuals registered in Registry Office lists are statistics units of the population of 'individuals' (1:1 ratio); (b) the administrative datasets used for the construction of the Statistical Register of Active Businesses (ASIA) in Italy detect the 'legal units' of enterprise that are sometimes in n:1 ratio with the statistical unit 'enterprise'; (c) in the EMENS dataset, produced monthly by the INPS (the National Social Security Institute), the statistical unit 'worker' is obtained by aggregating various 'contribution profiles' (1:n ratio).

Sometimes the reconstruction of the statistical unit is more complex and requires to be assisted by domain experts who are often the owner bodies of the source, the subjects of the administrative statement or the substitutes of the declarer in the communication to the body itself. An example of such a case can be found in the Archive of Companies listed on the stock exchange managed by CONSOB, currently used by the ISTAT for creation of the Statistical Register of Enterprise Groups: to identify the control links between enterprises starting from shareholdings between listed and unlisted companies, recourse is made to the knowledge and experience on the field of the owner body (CONSOB).

[2] Hence the term derived statistical objects.

Sometimes the statistical unit reconstruction can be assisted by the integration with other datasets, in the sense that some units require integration of administrative datasets to be identified. For example, (a) the reconstruction of households requires not only knowing the list of individuals and family relationship but also that they live in the same residential building, information derivable from a dataset about residences; (b) the reconstruction of the local unit also requires integration of the dataset of the Chambers of Commerce, which provides the addresses in which a business operates, with administrative/statistical datasets that provide information on the number of employees in the workplace in order to identify the statistical unit defined local unit.

The most frequent situation for the unit reconstruction is the mixed approach, which requires the joint aid of both the experts knowledge and integration with other datasets. One example is the statistical unit reconstruction "enterprise group" that, taking being an association of enterprises linked by decision-monitoring reports, on the one hand requires the help of experts of both economic, legal and tax (the accountants) and administrative (the owner body- Infocamere), to identify the control links among administrative units, and on the other side requires to know whether legal units are defined businesses, information derivable only by integration with the Statistical Register of Active Businesses (ASIA).

The complexity of reconstruction of the statistical unit starting with the administrative unit can be a function of the unit itself and grows if the statistical unit is complex, rather than simple, in the sense that it is an aggregation of base statistical units (of different nature) linked to each other by one or more constraints (relationships). One example is the statistical enterprise group unit that requires, in addition to base statistical units (enterprises), also control relations among enterprises. Of course, the composite statistical units are derived from the base statistical units.

In reconstructing the statistical unit, the so-called derivation error may occur, for the evaluation of which the tools available are limited. A measure can be provided by the number of units that cannot be univocally attributed to the population of interest. However, if the unit reconstruction is assisted by the integration, the derivation errors may result from *linkage* errors (see Principle B.2.), for the assessment of which it is often necessary to resort to manual control by experienced operators.

The units derivation process should be reproducible and documented. The assumptions underlying this process should be explained and documented.

Coverage evaluation. Once derived the target population, it is advisable to make an assessment as accurate as possible of the coverage error and therefore of the actual representativeness of the derived population compared to the target one.
In fact, the population of a single or integrated administrative source may differ from the statistical population of interest. This causes an erroneous coverage of the target population. A mismatch between the statistical population of interest and the one identified by the administrative source may depend on several factors:

(a) *conceptual difference between the target observation field and the one of the administrative source*. The conceptual difference between the two populations determines an inclusion of units not belonging to the statistical population (over-coverage) and a non-inclusion of units belonging to the population (under-coverage);
(b) *erroneous identification of the statistical units*. The complexity of the statistical unit reconstruction starting from the administrative unit can determine errors in the identification of the unit and therefore coverage errors. For example, if the administrative unit is in n:1 ratio with the statistical unit, the missing links in the identification of the unit generate an over-coverage error, while the erroneous links among administrative units cause under-coverage errors.

(c) *delays and/or missing administrative registrations*. Missing registrations generally determine under-coverage errors. However, when the administrative objects represent events (e.g. for certain demographic events), the missing registrations can also determine over-coverage errors caused by missing cancellations of statistical units from the reference population.

Where possible, several administrative sources should be used, because it helps to reduce both the under-coverage error, by integrating sources that cover different portions of the population, and the over-coverage error, having the possibility to dispose of more information to determine which units belong correctly to the desired field of observation. For example, as part of statistics on enterprises, the integration of different administrative datasets makes it possible to determine whether and which units belong to the observation field of active enterprises operating in economic sectors (ASIA register).

It is advisable to make an effort to estimate coverage error in the data. The ability to assess the coverage error is related to the availability of a *benchmark* to be used, possibly as a *gold standard* (statistical register, other administrative sources). There are two main approaches to measuring indicators on the coverage rate: aggregated comparisons with respect to known distributions and unit-by-unit comparisons (*matching case-by-case*). The first one provides coarser assessments. For an application, see U.S. Bureau of the Census (2011).The second one, more onerous because it requires activities of *record linkage* among datasets, is based on techniques known in the literature as Capture-Recapture (Wolter, 1986), or evolutions of these techniques that make it possible to relax some assumptions (Biemer P.P., 2011, pp. 249-258).

In many cases the *gold standard* of reference is not available and therefore it is necessary to find different methods of measurement. In the absence of a *gold standard* the classic method for the evaluation of the coverage of a list is the coverage survey, generally conducted for the evaluation of the coverage of census surveys.

Some bibliographic references

Biemer P.P. (2011). Latent Class Analysis of Survey Error. John Wiley & Sons, Inc., Hoboken, New Jersey.

Cerroni F, Morganti E. (2003). The methodology and informational potential of the archive regarding enterprise groups: first results. Istat Contributions 3/2003. http://www3.istat.it/dati/pubbsci/contributi/Contr_anno2003.htm

Cerroni, Di Bella, Galiè (2014). Evaluating administrative data quality as input of the statistical production process. Rivista di Statistica Ufficiale, no. 1-2, 2014

Blue-Ets (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

Eurostat (2010). Business Registers' Recommendations Manual

ESSNet Consistency (2013). Available at https://ec.europa.eu/eurostat/cros/content/consistency-0_en

Wallgren A. and B. Wallgren (2014). Register-based Statistics: Administrative Data for Statistical Purposes. Second edition: John Wiley & Sons, Chichester, UK. ISBN: ISBN 978-1-119-94213-9

US Bureau of the Census (2011). Source and Accuracy of Estimates for Income, Poverty, and Health Insurance Coverage in the United States: 2010 http://www.census.gov/hhes/www/p60_239sa.pdf

Viviano C., Garofalo G. (2000). The problem of links between legal units: statistical techniques for enterprise identification and the analysis of continuity. ISTAT. Rivista di Statistica Ufficiale 1/2000.

Wolter M.K. (1986). Some Coverage Error Models for Census Data. Journal of the American Statistical Association. Vol. 81, No. 394, pp. 338-346 ..

Zhang L.C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistics Neerlandica (2012), Vol. 66, no. 1, pp. 41-66.

Principle B.4. Variables derivation and classifications harmonisation

The variables derivation and classifications harmonisation process should follow established practices. All the assumptions underlying the derivation of the variables should be explicit and assessed their soundness. The validity of derived variables should be evaluated. The entire variables derivation and classifications harmonisation process should be documented.

Guidelines

After the statistical unit, the identification of the variables of the administrative source for statistical purposes represents the second step for the statistical use of the dataset. Similarly to what was found for the units, the variable derivation techniques are difficult to standardise and vary greatly among the different fields of application. Whenever possible, you should refer to the practices suggested in literature.

Even the classifications adequacy adopted in the administrative register should be assessed in the light of the necessary classifications in the processing and dissemination of the statistical variables. In this context, the "traceability" among classifications, namely the capacity to reconstruct each category of the statistical classification with a category of the administrative classification or with the joining of disjointed categories of the administrative classification, assumes particular importance. You should also have the definitions of the classification categories. In the case where the administrative data for some textual variables use a classification that cannot be traced back to the target one, it is advisable to have access to the textual data (not coded), which may enable full traceability to the classification of interest.

Classification errors, i.e. errors that are committed in aligning the classifications of the variables included in the administrative dataset with those of the statistical objective, especially if present in determinant variables for some statistical registers (geographical identifiers in population registers, industrial activity in the business registers) have impact on the identification of the populations of interest, and can in turn be the cause of coverage errors.

Identification and derivation of variables. Like the units, the statistical variables from an administrative dataset or from the integration of multiple sources are created by derivation from administrative variables (Wallgren & Wallgren, 2014) through a deterministic transformation/derivation function or through a random model. Both can be applied by using information internal to a single source or available from multiple administrative sources. In the latter case, the variables derivation activity involves a prior activity of integration among datasets, with possible errors that may be generated in the application of the *record linkage* procedures (see Principle B2 for more detail).
In general, errors that may be generated in the variables derivation process depend on an incorrect specification of the deterministic rules or of the random model. It is therefore advisable to make explicit the assumptions underlying the derivation of the variables and the harmonization of classifications and evaluate their validity.

The whole variables derivation and classifications harmonisation process should be reproducible and documented.

Assessment of the variables validity. The differences in concepts and in the data (the first represented by the specification error, the second by the measurement and processing error) between administrative and statistical variables should be explored and harmonised. The approaches that can be followed will vary depending on the availability or not of *benchmark* or control variables, and in particular depend on:

(a) the availability of direct control variables, namely statistical variables with coincident or linkable definitions from a survey or Census and which act as a *gold standard*, a situation which allows an exact

comparison between the data, the calculation of scale measurements, the calculation of shape parameters of error distributions and the calculation distance functions;

(b) the availability of "functional" control variables, i.e. not coincident from a conceptual point of view, but functionally linked with the variables of interest, which allows the application of *data mining* techniques (for *outliers* detection) and regression techniques, even multivariate for the study of functional relationships;

(c) the unavailability of either control variables or functional variables, which implies the use of approaches based on the study of consistency among variables (intra-source or among sources), factor analysis or latent class models that assume correlation structures among variables internal to the administrative dataset.

A generalizable example of validation of economic-accounting variables is traced in the work of Bernardi *et al.* (2013), which refers to the application carried out on the dataset of Sector Studies. It presents a validation methodology comprising the three type-situations described above with the relative quantitative validation methods associated.

It is appropriate to consider that the variables validation methods described should be applied also according to the type of use of the administrative data. In fact, if the aim is a direct use of the administrative dataset, it is desirable that the validation of the administrative source variable takes place through the use of control variables with the function of *gold standard*: the exact comparison with variables having coincident or connectable definitions ensures a high level of reliability in terms of microdata, a fundamental requirement when the goal is to replace the statistical variable with the one taken from an administrative source to produce direct statistics.
In the case of micro-integration (Bakker, 2010; Zhang, 2012), i.e. of integration in order to reconstruct the variables, if a *gold standard* is not available it will nonetheless be possible to perform a consistency check of information from multiple administrative or statistical sources, so a validation can be made by means of functional control variables. It should be pointed out that the use of functional control variables can be done with the aid of statistical models - algorithms for the micro-integration (Pannekoek, 2011) - or with the aid of experts. The activity of *profiling* performed on large enterprises in the Statistical Register of Active Businesses (ASIA) is an example of expert-assisted control (Eurostat, 2010).

Re-classification of the values of variables in derived units. The derivation of new units (Principle B3) naturally implies a variable coding problem related to the new units, which consists in calculating the value assumed by the variables, already present in the dataset, on the new units. One example is the recoding of all the variables referring to enterprises required during alignment between legal units and enterprises. Even in choosing the most appropriate method of classification or coding reference should be made to several possible methods, such as:

(a) the choice (deterministic or probabilistic) of the data contained in the administrative dataset considered more reliable in terms of the specific variable of interest (e.g. identification characters such as company name, address, etc.), frequent activity when integrating data from multiple administrative datasets;

(b) the assignment of a new value to the variable which is the result of a specific algorithm (e.g., the attribution of the main economic activity to the company by means of the criterion of the core business, or the calculation of employees in the local units contained in the Statistical Register of Local Units).

Some bibliographic references

Bakker B.F.M. (2010).Micro-integration: State of the Art. Note by Statistics Netherlands. UNECE Conference of European Statisticians. The Hague, The Netherlands, 10-11 May 2010

Bernardi A., Cerroni F. and De Giorgi V. (2013).A standardized scheme for the statistical processing of an administrative dataset. Istat Working Papers 4/2013

Eurostat (2010). Business Registers Recommendations Manual.

Pannekoek, J. (2011). Models and algorithms for micro-integration. Chapter 6. In Report on WP2: Methodological developments, ESSNET on Data Integration, available at https://ec.europa.eu/eurostat/cros/content/data-integration-finished_en

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. John Wiley & Sons, Chichester, UK. ISBN: ISBN 978-1-119-94213-9

ESSnetAdminData (2013). Final list of quality indicators and associated guidance. Deliverable 2011/6.5 of ESSnet on Admin Data https://ec.europa.eu/eurostat/cros/content/use-administrative-and-accounts-data-business-statistics_en

Zhang L-Chun (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica (2012) Vol 66, no.1, pp. 41-63.

Principle B.5. Time and territorial dimension

The different aspects of the time and territorial dimension should be carefully considered in light of the characteristics of the administrative data used, with respect to the rules of the administrative act on the territory, in relation to the life cycle of the administrative dataset.

Guidelines

An outstanding feature of the goal of interest in statistical production concerns the nature of the data to be produced: the estimate of interest can be cross-sectional, longitudinal in a microdata perspective (panel type surveys) or longitudinal in a macrodata perspective, i.e. for the production of time series (repeated surveys). Therefore defining the statistical product of interest implies identifying with precision the time and territorial dimension of the data used for this purpose. This means attributing the characteristics "place" and "time" to the definition of the population of interest, and consequently its constituent units. Similarly, each variable is associated with a time reference[3], be it an instant or a period of time. The variables of statistical interest can have a nature of stock variables or flow variables (Wallgren & Wallgren, 2014):

- the stock variables show the situation at a specific point of time, e.g. the age of an individual on a certain date, the number of employees of the firm at the end of the year;
- the flow variables represent sums in a given period, e.g. income received in one year, orders made by a company in a month.

By using the data of the administrative source, a correct identification of the time and territorial reference of the data is necessary, whether the data in question is used for the direct production of statistics or for the realisation of integrated microdata registers. Moreover, when it is necessary to transform the data to refer them to the desired time and territorial reference, it is important to verify the hypotheses.

With regards to the aspect relative to the time dimension we can identify datasets containing:

- stock data referring to a particular instant in time, updated in real time or on a date subsequent to the reference date;
- longitudinal information about the unit objects and events on births/deaths in time and their characteristics. Not all characteristics may be traced longitudinally.

The datasets may contain "time reference variables", namely variables that identify the time instant related to an event that occurs for an object, and in particular for a unit. For example, the "individual" unit and the "hospital admission" event can be associated with the time reference variable "hospital admission date." The format of these dates varies as a function of administrative procedures and can be identified by an exact date, as for a characteristic event of an individual of the population or by a period, e.g. a month for an event related to an enterprise.

It is evident that the time characteristics of the data contained in the administrative dataset condition the possibilities of statistical production, in cases where for longitudinal analysis requires longitudinal data, whereas to produce point estimates, *stock* administrative data are sufficient (these estimates can also be derived from archives with longitudinal data).

Concerning the geographical dimension, it is necessary a careful evaluation of the territorial reference of the administrative data used, because some administrative datasets may have a geographical coverage different

[3] A statistical variable is defined by: *i)*the unit that possesses the characteristic (eg income for people and income for households are two different variables) *ii)* the method of measurement, *iii)*the measurement scale and *iv)* the determined instant or period of time referred to the measurement (Wallgren & Wallgren, 2014, Chapter 8, pg. 148).

from that objective of the statistics to produce. This may require an integration of data from administrative sources and/or data from surveys.

The construction of statistical population or business registers, i.e. ideally complete datasets with respect to the above-mentioned populations, requires information necessary to trace and follow the units, their characteristics and relationships over time, but also information about the dates of the occurrences that arise in a given period of time. Since tracking objects and their identities over time is a costly business, it is appropriate to distinguish significant events for the purposes of statistical objectives - for which it is essential to know the instant of the occurrence or duration of the event - from the insignificant ones.

Just as it is important to know the date of occurrence of the events recorded in the administrative dataset, it is important to know their registration dates. The knowledge of the time references allows the construction of statistical registers that reflect the status of the population at a given instant or period of time. It is important to conduct scenario analyses to evaluate how the registration time of the events and the acquisition of administrative data affects other dimensions of quality, such as coverage of the target population or accuracy of the frequency estimates.

In the case of integrated data, the sources involved in the integration process may contain different time references, representing a further element of complexity associated with integration. The timeliness with which the longitudinal variations in the data are recorded in administrative datasets affects the quality of the integration activities and, consequently, can lead to other errors in the statistical data produced. The choices and assumptions used when integrating data with different time references must be properly justified and documented. The time frame of the output data resulting from the integration process should be clear.

In the datasets that contain longitudinal information, most of the objects do not present difficulties in the estimation phase, being units that are present for the whole reference period of interest, e.g. for the entire calendar year. Sometimes, the objects move in and out and may require the use of appropriate weights in the estimation phase. The time can be used as a variable that generates the weights, and used to correct the estimates.

The legislative and procedural changes, e.g. in the case of a change of classification adopted, may impact the comparability of the time series produced, and in particular may lead to changes in the levels of the series that do not reflect the actual context of the phenomenon. It is advisable to monitor these changes, whether they are documented in the metadata supporting the administration data (even late in relation to the release of the statistical data) or known in advance, in order to understand the real nature of the *shift* in the time series and act accordingly.

The procedures and techniques used to combine data from administrative sources characterised by different time references and/or different from the time references of the estimates or microdata to produce, and the weighting systems based on variables related to the time dimension, should be based on established methodologies and be reproducible and appropriately documented.

Some bibliographic references

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. John Wiley & Sons, Chichester, UK.

Principle B.6. Editing and imputation

The strategy adopted for the editing and imputation phase, when the data come from administrative sources, should take into account the peculiarities of the types of error of such data. The impact of the procedure must be assessed through appropriate indicators.

Guidelines

In the design of an editing and imputation plan (hereinafter E&I) the objectives of this phase should be given due consideration, which should not be seen simply as a data 'cleaning process', but as an unquestionable validation step.

The main objectives of a E&I procedure of the data are summarised below:

- identify possible error sources in order to improve the statistical production process;
- provide information on the quality of data collected and released;
- detect and correct influential errors;
- provide complete and consistent data.

These objectives are also valid in the context of E&I of data from administrative source.

The guidelines already developed for a classical statistical production process direct survey-type generally remain valid in this context[4]. In the case of data from administrative sources it is, however, necessary to take account of additional specificities characterising such data and having an impact on the organisation of a E&I plan.

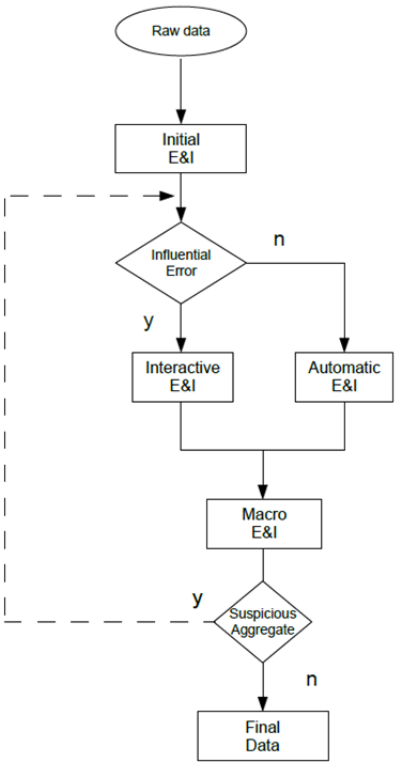
In the design of a E&I procedure for administrative data the first element to be considered is whether the statistics being estimated are obtainable using a single source or are derived from the integration of multiple sources.

Editing and imputation strategy in the case of single source of administrative data. When using a single source, the situation is similar to that in which it is necessary to design a E&I procedure of a single statistical survey, which typically can be represented by the following flowchart (Figure 1) drawn from the manual developed under the Edimbus project (Luzi *et al*., 2008).

In the first phase (the Initial E&I) all the evident and deterministic errors are corrected. Subsequently, the data set is divided into two subsets: one generally has few units, characterised by errors that are potentially influential and for which an operation of very accurate revision (Interactive E&I) is required, and another consisting of many units potentially affected by less important errors and that can be treated with automatic methods, which are generally more efficient in terms of costs (Automatic E&I). Finally, the estimates are produced and the initial data set is reassembled and controlled for suspicions of the estimates and compared with expected values or aggregates available from other surveys.

[4] Guidelines for the quality of statistical processes (http://www.istat.it/en/files/2011/11/QualityGuidelines_EngVers_1.11.pdf)

Figure 1. Flowchart of the operation of a data editing and imputation process



Editing and imputation strategy in the case of multiple administrative data sources. In the case of integration of multiple sources, it is necessary to assess at what point in the process to apply the E&I procedures. In general, there will be two alternatives for the overall strategy of editing and imputation, which constitute the two following scenarios:

1. E&I of each individual source→ Integration among sources → E&I of integrated sources
2. Integration of sources→ E&I of integrated sources

An advantage of the first scenario compared to the second is that it allows removal of some types of errors from the individual sources before the integration step (such as systematic errors due to errors of measurement units, balancing errors, etc.), thus reducing the possibility of consistency errors on the integrated data. On the other hand, in the first scenario, not all the available information is used jointly: e.g., for the imputation of a variable in a source, it may be useful to exploit variables observed in one or more of the other sources and closely linked to the variable to be imputed. Moreover, in the post-integration E&I it may also be necessary to remove any inter-source inconsistencies, generated by the very process of integration. The first scenario also involves high costs in terms of time and resources. A reduction in times and costs of this solution can be achieved by minimising the resources spent for the editing of individual sources, while ensuring an acceptable level of microdata quality: a solution in this direction could consist in performing on the individual sources only E&I of systematic errors and influential errors.

The second scenario, compared to the first, is characterised by a lower commitment of resources, since only one design of a E&I procedure is required. However, the procedure itself can be much more complex than that of the previous case. Furthermore, since the integrated data usually does not contain all the variables available in the individual administrative sources, some relationships between existing variables in the original sources may get lost.

In general, the choice of the more appropriate scenario is based on the trade-off between the expected quality level in the final data and resources actually available to achieve this level.

An element which in any case is of crucial importance for the purpose of increasing the effectiveness of the E&I process is the availability of administrative forms experts, who are familiar with administrative processes that have "generated" the data and their specific informational content, and who keep a close and continuous relationship with the bodies that supply the data.

It is evident that the use of longitudinal information can increase the efficiency of the overall E&I strategy.

Error types and treatment methods. In applications on data from an administrative source, the most significant and insidious type of error is the specification error, namely the non-correspondence between the statistical target definitions (reference population, variables) and those used for the production of the administrative data. This type of error is reflected in the variables observed in what is known as measurement error, and in particular a measurement error with a strong systematic nature. The assessment of whether or not this type of error is present in the data, which requires a thorough study of the concepts and the involvement of sector experts, is the first crucial step in any E&I strategy on data from an administrative source. For its detection there may be useful techniques for the detection of abnormal data applied to the difference between the administrative data and survey data (where available). For example, although the definitions of statistical target variables in administrative sources and survey have been harmonised, it may happen that for some subsets of data that harmonisation is not enough: this can give rise to a subset of data characterised by a large difference (relatively to the distribution of the differences among values) that may then suggest the presence of a problem on this set of data. For the correction of this type of error, deterministic approaches are applied, methods typically adopted in the case of systematic errors.

With regard to the localisation and treatment of the influential errors, the application of selective editing methods, as well as interactive editing in the administrative source data is limited by the dimensions of data and the difficulty of re-contacting the "primary source" (the respondent who provided the data to the owner body). On the other hand, there not being a sample selection process for this type of data, the selection of the units to be re-contacted is simplified by the fact that it must not consider the final weights, as is done in the classic surveys, and thus can simply be based on the identification of the most influential units, i.e. those generally having greater impact on the final results. It is in any case important to involve experts of the used administrative data in the process of interactive control of influential values, in order to maximise the possibility to correctly identify the cause of the error and its proper treatment.

As regards the use of automatic E&I procedures, e.g. those based on the principle of minimal change of the data, it poses no particular issues in the case of data from a single administrative source. In the case of data coming from a process of integration, where the same variable is available from several sources and inter-sources consistency errors can be generated, for the definition of the automatic procedure it is necessary to distinguish between two situations:

- case in which a source is considered as a *gold standard*, therefore not affected by error.
- case in which all sources are considered equally reliable.

In the first case through automatic procedures the validity of the information considered *gold standard* relative to its consistency within the source considered more reliable, is verified. In the case where an inconsistency occurs, e.g. due to a violation of a editing rule (*edit*), the data of the source considered less reliable can be used to reconstruct a consistent data item. An example is when the *gold standard* data is affected by a measurement error; in this case the data of the considered auxiliary source can help to unravel this type of error.

Where there is no source considered more reliable, all the observations can contribute to the reconstruction of a final value consistent with the expected constraints. The procedures used for the reconstruction of variables within the context of the data integration go under the name of *microintegration*. Some automated methods to treat the two situations described above can be found in Pannekoek (2014).

As well as in the survey data, the techniques of *macroediting*, i.e. the methods for identifying errors starting from comparisons of aggregates, may be useful for detecting errors in the data from a single source or in the integrated data.

Missing data treatment generally consists in the imputing namely in the "prediction" of the value which is not available or not usable in the administrative source. As regards the treatment of missing data it is necessary to emphasise some important aspects.

An essential element to be taken into account is that most of imputing techniques are based on the assumption that the probability of a non-response to a given variable is correlated with the observed values and does not depend on factors that cause the missing response itself (called MAR, *missing at random*). In essence this means that the available observations are sufficient to estimate a prediction model of the missing data, in terms less stringent the observed population is representative of that not observed.

With this premise, the first item to be considered is therefore the cause of the missing data.

The data may be missing within a source because the variable in question is not of main interest to the administrative body that collects the information, in this case the subject of the administrative act may not be sufficiently solicited to provide information on this variable. This situation is similar to that which occurs in surveys for which a MAR mechanism is frequently speculated.

Another important case found in the use of administrative sources data is what you get when you combine different sources that do not observe all the same variables. In this case there is a missing response that cannot be considered random, because it corresponds to the different segments of the population identified by the different sources. Hypothesising a MAR mechanism for such situations therefore corresponds to accept the idea that the population segment in which the variables are observed have a similar behaviour in terms of structure of the variables with the missing response to segment of the population not observed.

For example, suppose you want to estimate the income statement items of all Italian businesses using the source Statutory Accounts (SA) and the Tax returns source (Unico form). For the subpopulation of Joint-Stock Company, the SA source provides all the information necessary to estimate the detailed income statement items. For the subpopulation of Professionals, it is not possible to use the SA source, the Unico Source provides information useful to the estimate of a subset of items in the income statement. In estimating other goal variables, the information of the SA source (only information available) must be used and this implies the hypothesis that Professionals and Joint-Stock Companies are characterised by homogeneous conduct in terms of phenomena subjected to estimates.

As already mentioned, when using administrative data, since often the individual sources are not exhaustive of the population in question, the joint use of several sources is frequent and therefore all the sources must be integrated.

The integration of sources involves the possibility of introducing additional types of errors. In integration procedures there are *false links* and *false non links*. The errors introduced by *false non links* cannot be treated in the editing and imputation phase because they do not give rise to observations to control. The *false links*, on the contrary, generate units whose information content (e.g. the variables from the various sources) is inconsistent. The inconsistency among the variables of a unit can also be found in cases where the combination was effective, in fact, although it is assumed that the sources have been harmonised, a

misalignment is always possible, even if mild. We can summarise by saying that specific errors in this context and not present in classical surveys cause intra-source inconsistencies, which are inter-source inconsistencies naturally present or artificially introduced by a false match. There are various techniques to handle this type of errors that are aimed at the reconciliation of data, also in this case reference is made to the so-called *microintegration* for which important references can be found in https://ec.europa.eu/eurostat/cros/content/data-integration-finished_en and in Pannekoek (2014).

Assessment of the procedure. Regardless of the strategy adopted, as an integral part of the E&I procedure, there should be a calculation and monitoring of the sets of quality indicators on the input and output data, such as frequency of activation of control rules, imputation rates, indicators of impact on distributions (Luzi *et al*., 2008). What is more, since the administrative data collection process is outside the control of the Statistical Institutes, it is important the identification of instruments for signalling possible changes in the data formation process at the supplying institution, for the monitoring and reduction of the effects of these changes on the released statistics.

Some bibliographic references

de Waal, T.,Pannekoek, J., Scholtus, S. (2011) *Handbook of Statistical Data Editing and Imputation*, Wiley.

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Tempelman C., Hulliger B., Kilchmann D. (2008) Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. EDIMBUS project https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Pannekoek J., (2014). Method: Reconciling Conflicting Micro-Data, in the Handbook for Modern Business Methodology. http://www.cros-portal.eu/content/reconciling-conflicting-micro-data

Principle B.7. Estimation process

For the purpose of estimates' production, the data acquired from administrative sources and properly treated should be processed according to methodologies, which take into account the specificity of the informative and productive context (presence or absence of survey data). When using statistical estimation models, the assumptions underlying these models should be properly spelled out and their actual validity has to be checked. The estimates produced should be accompanied by error estimates, in order to allow proper use and interpretation of the results.

Guidelines

The procedure for deriving the estimate of interest (estimates of levels, ratios, contingency tables, etc.) using administrative sources should be clear, well defined and possibly rely on consolidated statistical techniques.

The estimation process by means of the administrative data (database in which the various available sources have been integrated) mainly depends on the following assumptions:

(1) the statistical target population is covered by the administrative source(s) used;
(2) the administrative variable observed coincides in the definition with the variable of interest and differences, if any, are random and not systematic;

and from the following informative scenarios:

(a) there are only the administrative data and the size of the reference population and possibly the size of subpopulations of interest are known;
(b) there is an unbiased estimate of the parameter to be estimated with the administrative data;
(c) there are reliable estimates of parameters correlated with the parameter of interest.

The approaches used depend on the assumptions and the scenarios listed above, but also vary as a function of the direct or auxiliary use of the administrative data.

Direct use of administrative data to estimate. If assumptions (1) and (2) are valid, i.e. if the administrative data covers the whole population, the estimation of a total is obtained as a simple sum of values.

In general, however, the assumption (1) is rarely satisfied and the administrative data offers partial coverage of the population of interest, thus representing a non-probabilistic sample of the population itself. In this case, a treatment of the missing values is recommended and should be appropriate and coordinated with respect to the variables to consider (Wallgren and Wallgren, 2014). In general, when the condition (2) holds by partial coverage of the interest population, the process can make use of the two following estimation approaches:

- a statistical prediction model should be clearly explicited and should be based on reasonable and possibly verifiable assumptions. The main objective of the statistical model is the prediction of the values of the variable of interest in the part not covered by the integrated administrative sources. The statistical model uses as covariates the variables known in the covered part of the population but which are not covered by the administrative data. This approach, in general, is appropriate when integrating data from different registers to cover the population and missing values occur for a limited set of variables of interest;
- a calibration method that calculates a weight for each unit that presents the administrative data according to the logic of the approach to calibration that leads to known totals for the entire reference population (Wallgren and Wallgren, 2014). The process implicitly considers a statistical model in which the covariates are represented by the same variables that define the known totals. This approach is most appropriate when a large number of administrative source variables for the unit are not available.

In both estimation methods it is vital the hypothesis that the estimated model in the part covered is valid for the part not covered. This hypothesis is verified only approximately, and thus can cause bias in the estimates, the impact of which can nonetheless be reduced in terms of its relative MSE.

Establishing a parallelism with the theory of the finite population estimate, these two procedures represent an imputation process and a calibration process. In the first case a random imputation (by perturbation of the predicted value), or a deterministic imputation can be used. Only in the case of variables with a very low rate of missing values and evenly distributed in the domains, that is, in sub-populations of interest, it can be assumed to produce estimates by sum, indicating the rate of missing values.

In the type (a) informative setting indicated above, the two approaches to the estimate use only the information from the administrative dataset. In particular, for the estimation process which uses the second method (calibration method), it is appropriate that the weights sum up to the known population/subpopulation sizes.

When there is an informative scenario (b) various uses of administrative data are possible in the estimation process:

- one can estimate the statistical model using sample units not covered by the administrative sources (Latilla and Holmberg, 2010). The MSE of the sample estimates and the estimates from the administrative data based on the statistical/calibration model are compare and the estimator with the lowest MSE for domains with more details of interest is chosen;
- the values of the variable of interest in the part not covered by the administrative data are predicted by means of a statistical model and a compound estimator, that is given by the convex combination[5] between the sample estimate and that of the administrative data, is defined. The weights of the two estimates are inversely proportional to their MSE (Moore *et al*., 2008). This approach can lead to mixed estimation processes in which part of the estimates is completely sample-based and part is totally from administrative data;
- the administrative sources and sample data are combined at unit level through integration techniques. A hierarchy is defined on the reliability of the value observed between administrative data and sample-based data. The universe is reconstructed by integrating the two sources and choosing, in the event of overlap, the hierarchically more reliable value between the administrative data and sample data. The total of the not covered part is estimated by the sample estimated (Kuijvenhoven and Scholtus, 2010, 2011).

In the three estimation processes it is assumed that the size of populations/subpopulations of interest are known.

In the case in which the informative setting is of type (c), the estimation process can be entrusted to a calibration method that exploits the values of the estimates correlated to the parameter of interest.

Auxiliary use of administrative data in the estimation process. When condition (1) of the initial assumptions is not valid, that is, the administrative data item does not exactly reflect the variable that defines the parameter of interest (systematic error), but condition (b) is met, then the administrative data can be used as an auxiliary source. In particular, this strategy improves the efficiency of sample estimates if there is a correlation between the variable in the administrative register and the variable of interest. The estimate in this case should make use of statistical models or of the calibration method that use the final direct weights:

- the sample estimation process is refined by using the administrative data. In particular, the informative patterns homogeneous with respect to the variables available from the administrative sources are identified. The sample estimate is then estimated by the calibration estimator for each subpopulation that presents a given pattern of administrative variables. The final weight unique for each unit of the sample;

[5] Convex combination: linear combination of estimators with non-negative coefficients whose sum is 1.

- specific sampling estimators can be used for each variable of interest. In this case a projection estimator (Luzi *et al*., 2014), which uses a statistical model for each variable of interest, is defined. The specific model presents as covariates the ones most closely correlated with the variable of interest available in the administrative variables pattern. The predicted values impute the missing data. The model is estimated with the weights of the probability sample. A massive imputation of the entire reference population for each variable of interest is produced (Kim and Rao, 2011).

Regardless of the specific use of administrative data, where possible, an effort should be made to estimate the unbiasedness and accuracy of the results, using advanced methods, including Bayesian methods and Structural Equation Models ((Bryant J.R. e Graham P.J., 2013; Scholtus S. and Bakker B.F.M., 2013), as long as they are applied to a context of validity of the assumptions that allow their application. In any case, the estimation process and the assumptions on which it is based should be appropriately documented, and an assessment on the possible errors generated by this phase should be attempted.

Some bibliographic references

Bryant J.R. and Graham P.J. (2013). Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources. Bayesian Analysis (2013), 8, no. 3, pp. 591-622.

ESSnets Admin Date (2013 Guidance on the Accuracy of Mixed-Source Statistics). Deliverable 6.3/2011 USE OF ADMINISTRATIVE AND ACCOUNTS DATA IN BUSINESS STATISTICS March, 2013.

Kim, J. K. K., Rao, J. N. K. (2011). Combining data from two independent surveys: a model-assisted approach. Biometrics No. 8, pp. 1-16.

Kuijvenhoven, L. and Scholtus S. (2010). Estimating accuracy for statistics based on register and survey data. Discussion Paper 10007. Statistics Netherlands, The Hague/Heerlen.

Kuijvenhoven, L. and Scholtus S. (2011). Bootstrapping combined estimator based on register and sample survey data. Discussion paper 201123. Statistics Netherlands, The Hague/Heerlen.

Laitila, T. and Holmberg, A. 2010. Comparison of sample and register survey estimators by means of MSE decomposition. Paper for the European Conference on Quality in Official Statistics, May 4-6, Helsinki.

Luzi, O., Guarnera, U., Righi, P. (2014). The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data. European Conference on Quality in Official Statistics (Q2014). Vienna, Austria, 2-5 June 2014

Moore, K., Brown G. and Buccellato T., 2008. Combining sources: a reprise. Paper for the CENEXISAD workshop "Combination of surveys and admin data, 29-30 May, Vienna.Office for National Statistics, Annual Business Survey. http://www.ons.gov.uk/ons/guidemethod/method-quality/specific/business-and-energy/annual-business-survey/index.html(accessed on 12/08/12).

Scholtus S. and Bakker B.F.M. (2013). Estimating the validity of administrative and survey variables through structural equation modeling. A simulation study on robustness. Discussion Paper (2013)

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. Second edition: John Wiley & Sons, Chichester, UK.

Principle B.8. Validation of the results

The analysis results, prior to publication, should be evaluated together with sectorial experts to check whether or not there are anomalies. Whenever possible, the results should be compared with those obtained in previous replications of the same process, where data from the same administrative source were used. Alternatively, the comparison can be made with similar results obtained by other processes of the same institution or of other bodies. In addition, process quality indicators should be calculated and analysed rigorously.

Guidelines

Before being disseminated, the results of the statistical production process, even when data from administrative source have been used, should be evaluated by comparing them with the results of previous editions, where the data from the same administrative source were used and by comparing them with statistical sources internal or external to the Institute. Any differences found should be justified and documented.

If possible, consistency of the results should be controlled with respect to ratios that can be considered nearly constant or subject to minimal changes in the short term, such as certain demographic ratios. In this case too, any differences should be justified and documented.

Moreover, before the release of the data, in case of suspect values, the results should be assessed by experts of the Institute or external experts, primarily representatives of the institutions supplying the administrative data, but also possibly representatives of academia or trade associations. If the control is carried out by the experts external to the Institute the respect of data confidentiality should be guaranteed. In any case it is preferable to involve in the validation experts, internal or external, who are not directly engaged in the production of the data item.

In the validation phase, the quality indicators available on the input data and those calculated during the data treatment process (e.g. indicators of matching errors) should be systematically analysed and compared to the expected levels of these indicators or, in any case, with the aim to evaluate the points of weakness of the process and identify possible corrective actions.

The calculation and analysis of quality measures and process indicators, are designed, firstly, to ensure the quality of the estimates disseminated and, secondly, to assess the appropriateness of adopting improvement actions for subsequent editions of the process.

In cases where there are margins for improvement by adjusting the administrative source, the result of the assessment should be translate into feedback information for the body owner of the administrative data, through the centralised structure that handles relationships with the administrative body.

Principle B.9. Archiving, confidentiality protection, data dissemination and documentation

The validated micro-data, properly supported by metadata and quality measures, should be archived according to the Institute standards prior to its internal dissemination for further statistical uses and before its external dissemination. The disseminated macro and micro-data should be pre-treated to ensure adequate protection of confidentiality. The dissemination calendar of the statistical results should be made public. All the phases of the process should be properly documented.

Guidelines

The validated micro-data should be archived together with the metadata necessary for their interpretation (record layouts, variables and associated classifications) in the Institute systems[6], following the procedures defined by the Institute, also in the case in which administrative source data are used within the production process. Being the micro-data from administrative data frequently an intermediate product, used as input of other statistical production processes, it is vital to measure their quality by means of specific indicators, such as coverage, missing data, timeliness and punctuality.

The goal of dissemination is to enable a timely and effective use of the information produced by the Institute, thereby meeting the needs of users. For this purpose it is useful to define in advance a dissemination calendar for the various types of releases, which should be made public to users. Access to released data should be simultaneous for all users to ensure the impartiality and independence of the official statistics.

To allow a better use of data by users it is important to disseminate data that are easily accessible and understandable. Accessibility is linked to the type of medium used (online releases, CD-Rom, paper volume), and the ease of information retrieval. Given the current national and European guidelines, Internet has become the predominant dissemination mode, both through the implementation of data warehouse, and through the publication of documents, press releases and online volumes. The clarity, however, is linked to the availability of metadata related to the informative content and the characteristics of the production process, and the quality indicators. The support metadata should be integrated with the elements that allow us to understand how the administrative data were used and their validity in the specific production context. In addition, any data limitations, such as time series breaks and the possible provisional nature of the released data should also be communicated.

The various types of issue, e.g., press releases and yearbooks, should comply with publishing standards.

The law establishing the National Statistical System, Legislative Decree 322/89 requires that the confidentiality of the respondents must be protected, and, in particular, that the data being disseminated have to be adequately treated for that purpose. In the case of aggregated data published in tables methods can be used such as the threshold rule, which is set as equal to or greater than three, and the methods which consist in perturbation of the data so as to reduce the possibility of identification and acquisition of information on the individual units. In case of release of elementary data, specific methods can be used, such as recoding variables to reduce the informational detail, the removal of specific information that can make the unit identifiable, and methods of perturbation of the elementary data. For the protection of confidentiality in data disclosure it is recommended to use generalized software.

[6]For the data produced by the surveys, the Institute standards envisage storage in the ARMIDA repository (ARchive of validated MIcroDAta) which was founded with the primary goal of preserving and documenting the data, subsequently associated with the goal of disseminating the data itself. The data stored in ARMIDA feed, in fact, the different distribution channels of microdata (for internal use to the Institute through "Access memorandum to ARMIDA" microdata for internal users", for of Sistan institutions, for files used for the research, for the standard files, etc.). The micro-data stored in ARMIDA are also used to respond to requests of external users presented at the ADELE laboratory.

The production process should be properly documented, with regard to all phases, from the evaluation and choice of the administrative source, to the integration of data in the statistical production process, to the treatment of integrated data, until the final publication.

This documentation should include quality indicators, such as indicators of timeliness, coverage and missing data, consistency and comparability over time.

Some bibliographic references

Hundepol A., Domingo-Ferre J., Franconi L., Giessing S., Lenz R., Naylor J., Nordholt E.S., Seri G., De Wolf P.P. (2010). Handbook on Statistical Disclosure Control. Version 1.2. ESSnets SDC - A network of excellence in the European Statistical System in the fields of Statistical Disclosure Control http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf

Istat (2004). Metodologie e tecniche di tutela della riservatezza nel rilascio di informazione statistica. Metodi e Norme, n. 20 http://www3.istat.it/dati/catalogo/20040706_00/manuale-tutela_riservatezza.pdf

Istat (2008). Protocollo d'accesso ai microdati di Armida per gli utenti interni https://intranet.istat.it/MetadatiEQualita/Documents/Microdati%20validati/ProtocolloArmida.pdf

Istat (2009). Standard di documentazione e di memorizzazione dei file di microdati per la ricerca (MFR), Servizio SID, 26 giugno 2009 https://intranet.istat.it/Documentazione/Procedure/MFRStandard.pdf

Istat (2009). Procedura per il rilascio di file di dati elementari agli uffici Sistan, Ordine di servizio n. 148 del 17 novembre 2009 della Direzione Generale https://intranet.istat.it/Documentazione/Procedure/Procedurarilascio_sistan.pdf

OMB (2006). Standards and Guidelines for Statistical Surveys. Office for Management and Budget, The White House, Washington, USA http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/statpolicy/standards_stat_surveys.pdf

Part C. Output quality

1. Introduction

In parts A and B the principles and guidelines to be followed in the evaluation of administrative data used a input and in the conducting of a process that uses them, respectively, in order to produce statistics characterised by high quality in an efficient way, are listed. Nonetheless, having set up and conducted a high quality process does not imply that the quality of the statistics produced has not to be measured. As in the Guidelines for the statistical production processes of "direct survey-type", this section summarises the criteria against which to measure the quality of statistics produced and with respect to which communicate it to users, taking account of the specificities of data production, i.e. introducing elements related to the impact that the use of administrative source data can have in the measurement or in the levels of the measures of quality. However, genuine guidance on how to conduct the measurement, requiring a deep methodological study for which reference should be made to the specialised literature, is not provided.

2. Definition and dimensions of product quality

First, it is advisable to reiterate the meaning and the scope for which quality is defined. For the purposes of this discussion, product means the final product resulting from the statistical production process or output (a term often used even now in the Italian terminology). This product has a statistical nature (distributions, level estimates, variations, etc.). Products that could be called "intermediate", such as statistical registers, are not considered here.

For the purpose of measuring the quality of statistics, ISTAT adopted the definition of quality issued by Eurostat in 2003 (ESS Working Group "Assessment of Quality in Statistics"), later taken up by the European Statistics Code of Practice (2011) and the Italian Code of Official Statistics (Official Gazette no. 240, 13 October 2010)[7]. These definitions have been further clarified in reference Eurostat manuals for *quality reporting* (Eurostat, 2009; Eurostat, 2014).

The quality is defined as "the totality of features and characteristics of a product or service that bear on its ability to satisfy steted or implied needs" (Eurostat, 2002 Eurostat, 2003a). In this sense, the quality of the statistics produced and disseminated must be assessed with reference to the following criteria (Eurostat, 2003a):

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

Afterwards, the reliability was added to the accuracy component. In these guidelines the most updated Eurostat definitions of quality criteria have been reported. They are taken from the *ESS Handbook for quality reports* (Eurostat, 2014). In some cases the definitions have been reformulated and summarised without changing their meaning.

[7] This definition of quality has taken on considerable importance as it has been included in the legal framework (Regulation (EC) No. 223/2009 of the European Parliament and of the Council of 11 March 20097) which regulates the production of European statistics.

Definition C.1. Relevance

Relevance is an attribute of statistics measuring the degree to which statistical information meets current and potential needs of the users.

Definition C.2.1. Accuracy

The *accuracy* of statistical outputs in the general statistical sense is the degree of closeness of computations or estimates to the exact or true values that the statistics were intended to measure.

Definition C.2.2. Reliability

Reliability refers to the closeness of the initial estimated value to the subsequent estimated value.

Definition C.3. Timeliness and punctuality

Timeliness describes the length of time between results availability and the event or phenomenon they describe.

Punctuality is the time lag between the actual delivery of data and the target date on which they were scheduled for release as announced in an official release calendar, laid down by Regulations or previously agreed among partner..

Definition C.4. Coherence and comparability

Coherence measures the adequacy of the statistics to be combined in different ways and for various uses. These concepts are further broken down into: coherence across domain, i.e. the extent to which statistics are reconcilable with those obtained through other data sources or statistical domains; coherence between sub annual and annual statistics, i.e. the extent to which statistics of different frequencies are reconcilable; coherence with National Accounts, i.e. the extent to which statistics are reconcilable with National Accounts; internal coherence, i.e. the extent to which statistics are consistent within a given data set.

Comparability (geographical, over time) is a measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas or over time.

Definition C.5. Accessibility and clarity

Accessibility refers to the simplicity and ease, the conditions and modalities (e.g. where to go, how to order, delivery time, pricing policy, formats…) by which users can access data.

Clarity refers to the simplicity and ease by which users can use and interpret statistics, with the appropriate supporting information (e.g. metadata, illustrations, quality documentation, …) and the extent to which additional assistance is provided by the producer.

3. The measurement of the quality of the statistics produced using data from administrative source

Especially for statistics that use data from administrative source, measuring their quality according to the above mentioned components it is not at all easy. In fact, only some components lend themselves to a direct quantitative measurement, in particular timeliness and comparability in time, while for the other dimensions often only judgments can be formulated. With respect to accuracy, it must be emphasised that the sampling error component generally does not apply and the non-sampling error is characterised differently from the direct survey situation.

In the following the quality components and how the use of administrative data may alter their interpretation and the measurement of related indicators will be analysed. So, the main non-sampling errors that are generated in the process that uses administrative data will be listed and defined, as shown in Figure 2 of Section 1 "Framework for the quality of statistical processes using administrative data", useful as a reference for the identification of indirect measures of the accuracy of the statistical data derived from those sources.

3.1. Measure the quality components for processes that use data from administrative source

We can say that the use of administrative data does not alter the meaning of the quality components: the statistics produced will be subject to considerations on the relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, accessibility and clarity just like statistics from direct surveys. The use of administrative data for statistical purposes, however, particularly affects some of the quality components, and sometimes limits the ability of the researcher to evaluate them.

The Relevance and Coherence and Comparability can be strongly influenced by the use of administrative sources directly for statistical production (replacement of units and/or variables). In particular, following the model introduced in Section 1 "Framework for the quality of statistical processes using administrative data", specification errors, i.e. the mismatch between statistical goal concepts and the related concepts underlying administrative data, may introduce in the derived statistical product, *relevance* errors. Even where such errors do not occur, legislative or procedural changes, that impact administrative populations and variables, can result in a lack of *comparability* in the statistical data produced from administrative data. The integration processes among datasets may introduce consistency errors among sources, which are to be added to the possible problems of internal consistency within the sources, that it was not possible to reconcile through the proper harmonisation procedures and editing and imputation procedures.

The timeliness and punctuality indicators remain perfectly valid in the case of statistics produced by using administrative data. The use of administrative data may impact on *timeliness* in both directions: where the availability and supply of administrative data fits the needs of statistical production, its use can lead to gains in timeliness. Conversely, if the administrative data is available with delay in relation to production requirements, this may produce a worse timeliness of the statistics produced, with respect to the hypothetical situation of a direct survey.

Considering the *accessibility* and *clarity*, it can be said that the use of administrative data has no impact on the measurement of these components. To ensure transparency to users, it will be appropriate to document which administrative sources have been used and how.

Surely, the component that presents greater difficulties in measuring direct surveys as well as in processes that use administrative data is that of accuracy. Often the administrative data are used in combination with those measured by direct surveys, and therefore even the estimators used may have different forms, ranging from situations in which data are integrated at the micro level up to cases in which the estimator is a weighted average from estimators from survey and estimators from administrative data.

3.1.1 Accuracy and reliability

The level of accuracy of the results is related to the number of errors that may occur in the production process of the estimates. A measure of accuracy is provided by the *Mean Square Error* (MSE), which includes variability and bias for all sample and non-sample components (the former not applicable or not relevant in the context of use of administrative data).

For statistics that use administrative data, not many experiences of data accuracy estimate are available in the literature. Laitila and Holmberg (2010) propose a breakdown of Mean Square Error which also takes into account the bias from relevance, beyond the classic ones of the estimator, from non-response and measurement error. Under certain assumptions, they provide a method of comparing the accuracy of the register estimates with those from sample survey and show that the worst distortion of the estimator attributable to the error of relevance can be compensated by the absence of sampling variability.

The work conducted within the WP 6 of the ESSnet "Use of Administrative and Accounts Data in Business Statistics" (Deliverable 6.3., 2011) focused on the identification of accuracy measures for "mixed" sources in different situations of micro and macro integration appears to be interestingly. Estimates of variance and bias through methods of "bootstrap re-sampling" are also possible. Estimators and associated credibility intervals can be derived through the use of Bayesian methods.

Interesting, albeit not yet applicable to regime, the use of Structural Equations Models to estimate the validity of the administrative and survey variables (Scholtus S., Bakker B.F.M., 2013).

Statistics produced by using administrative data are potentially evaluated by calculating reliability measures, i.e. through analyses of revisions. Please remember that the revisions are regularly scheduled and successive updates of an estimate due to update factors (using more updated sources, adjustments for seasonality) or improvement (in definitions, methods). So when administrative data becomes available with varying levels of completeness and updating, the quality of the statistics produced by the different datasets can be evaluated through revision indicators. See the website of the OECD and the ONS for a comprehensive treatment of the subject.

As done for the survey-type statistics, in conditions of difficulty in deriving reliable and low-cost measures of the mean square error, often an assessment of the quality of the results is carried out by analysing the different error components that impact the accuracy of the estimates. Here below are the descriptions of the major errors in the case of statistics produced by using administrative data, always with reference to Figure 2 of Section 1 "Quality framework for the quality of statistical processes using administrative data."

3.1.2. Errors on units or population

Lack of conceptual correspondence between the statistical population and administrative population

If at a definition level, the units that make up the statistical target population do not correspond to those underlying the administrative dataset that should contain them, this mismatch can lead to an undercoverage or an overcoverage. However, the problems of coverage can become depleted not only in this phase, but may originate from the process of derivation of the statistical population starting with the administrative population through procedures of integration, derivation, recoding. In this case the coverage will also include other components that may be derived from errors during these phases (see below). Suppose we have a statistical target population of "building permits" defined as: projects for new buildings (residential and commercial), expansions of existing buildings, executive projects for the construction of buildings or extensions intended for public housing. If the reference population of the administrative source available was represented by the building or expanding permits only for residential and non-residential buildings, a

coverage error would be present. In the terminology used in Zhang (2012), the set of units regarding public construction projects "is not accessible."

Selection errors

The selection errors originate from the discrepancy between the accessible (or observable) set of data of interest and that concretely acquired. Returning to the hypothetical example on building permits, let's assume that all building permits are submitted to municipalities, some in paper form (residential and non residential) others in electronic form (public housing) and are then transmitted by the municipalities to the ISTAT. If the electronic transmission system suffers a malfunction, and then a part of the permits fails to get sent to the ISTAT, we are in the presence of a selection error. Again, in the efficacious terms of Zhang (2012), all the units regarding public construction projects would be "accessible but not accessed".

Linkage errors or matching error

In integrating different administrative datasets between them and with survey data, the main errors that are committed are *i)* false non-matches and *ii)* false matches. It is evident that the quality of the results of the match depends strongly on the quality of the matching keys used. The matching errors have an impact on other components of the error. Mainly, the false non-matches may lead to coverage errors and missing data errors, while the false matches may lead to consistency errors in variables.

Unit derivation errors

This category includes various types of errors that may generate during the processes of derivation of the unit. In particular, errors in the *ex-novo* creation of the unit, i.e. unit that does not exit as such in the administrative sources, e.g. it is the case of the statistical unit “agricultural holding” for which it is necessary to look for existence signals in the various administrative sources. In addition, errors in the identification of the statistical unit starting from the administrative one are included. Finally, errors from aligning composite and “base” units are part of this category. For example, if you have access to data at the individual level in different data sets and want to form a list of resident individuals by residence, if the different data sets contain different addresses, condition that must lead to a decision on the address to assign to each individual, errors can be committed in the identification of the resident individual.

Coverage errors

The coverage errors are errors arising from a discrepancy between the statistical target population and the one derived in the production process that uses administrative data. The latter, in the simplest situation, can coincide with the reference unit of an administrative dataset, or it can be the result of integration and derivation procedures of units that can be very complex.

Within the coverage error the usual categories of errors can be identified: *i)* undercoverage, or objects that belong to the target population but are not listed in the dataset (or in the integrated datasets), which represent a potential source of bias; *ii)*overcoverage, namely objects in the dataset (or in the integrated datasets) but not belonging to the target population of objects that represent a potential source of variability; *iii)* errors in identification variables of the objects, which may lead to subsequent integration and consistency errors.

When using administrative data, the coverage is the result of a set of possible components, such as: the dataset coverage compared to the administrative population (which sometimes refers to administrative objects of the event type), the coverage between statistical target population and populations potentially detected from the administrative dataset, coverage between the population detected from the dataset and acquired population (selection error), the statistical target population and the one derived from the process of integration and derivation. In fact, bear in mind that linkage errors in the integration procedure, and in

particular false non-matches can cause undercoverage errors. Also possible delays in the acquisition of the data and the unit identification error can lead to coverage errors.

3.1.3. Errors on variables

Specification errors

Just as in the case of direct surveys, the specification errors result from a mismatch between the cognitive goals of the survey and relative concepts observed through the questions in the questionnaire, in the use of administrative data, these errors are related to discrepancies between the statistical theoretical target concept and the administrative one. It is perhaps the type of error that has the most impact on the relevance of the statistics produced, and can cause accuracy errors, particularly in terms of the bias.

Lack of conceptual stability

Legislative or procedural changes that govern the administrative act and amend the concepts underlying the administrative data can lead to errors of comparability over time. Lack of regional legislation homogeneity lack of uniformity in the procedures for acquiring and processing of administrative data can lead to errors of geographical comparability. The time and geographical comparability errors are propagated in the statistics produced by administrative data; however, at the origin there is not necessarily an error in the sources used, but simply some variations (time and geographical).

Measurement errors

These are observation errors that may occur in the collection phase (*measuring errors in the strict sense of the term*). In practice, the value available for a given variable, at the time of dataset acquisition does not correspond to the true value. Such errors may be both a source of bias and an increase of the variability associated with the estimates.

Processing errors

These are errors generated in the treatment of the data within a given statistical production process (revision, data entry, coding, control, processing, etc.). Some of these errors may also result from treatments previous to use, conducted by the owner body or, more rarely, by the centralised sector of data acquisition. In fact, the pre-treatment in the phase of centralised procurement, generally does not change the data acquired, but supplements them with additional information.

Misclassification errors

These are the errors committed when aligning the classifications used in the administrative data with those intended to be applied to the statistical variables of interest, namely in assigning to each category of the administrative classification a category of the statistical classification. Misclassification errors in stratification variables can lead to coverage problems.

Consistency errors

The consistency errors are derived from the violation of a series of compatibility rules within a single source or among multiple integrated sources. In the first case we speak about intra-source, in the second about inter-source consistency. The latter may or may not result from false matches in the integration procedure.

Errors of missing data

These errors are similar to those which in the case of survey statistics are referred to item nonresponses, whereas the so called unit nonresponses correspond -- for administrative data -- to coverage errors. Typically

the non-observation of data is observed in full for some variables that are not of interest for administrative purposes, while it does not appear to be relevant to the variables of strictly administrative interest. This type of error can result even when variables present in different datasets are integrated, as the effect of the integration process among datasets, where the item that is integrated is not present in all the sources. The impact on the estimates of this type of error is in terms of variability increase and possible bias.

Specification error by implicit or explicit model

The model assumption errors occur every time a model is introduced, in general for adjustments (missing data, seasonal effects) and for estimation operations. Statistics produced by using administrative data typically make extensive use of model-based approaches during the activities of integration, derivation of units and variables, editing and imputation, estimate. Typically a model is a set of assumptions about relationships between observed data and unobserved data. The validity of the model assumptions should be checked even if there is no access to tools with which to evaluate it with certainty. To assess the impact of the model assumptions, it is possible to resort to a sensitivity analysis through simulations.

3.2. Quality indicators

The difficulties involved in measuring accuracy or reliability and, more generally, the single quality components makes the most widely used approach to the measurement of quality consists in a compromise: the few direct measurements are put beside the indirect ones. In the case of the use of administrative data, these measurements can be defined both on the input data and for the integration phases of the data in specific statistical production processes. The input data indicators, placed beside other more contextual clues about the source, provide a concrete framework on the usability of administrative data for statistical purposes.

On the input data, the most significant literature is made up of the evidence produced in the framework of the FP7 project "Blue - Enterprise and Trade Statistics (Blu-Ets)", which is consistent with the approach that defines the input quality through three main hyper-dimensions (source, metadata and data), developed for the hyper-dimension "data" and for its different dimensions, quality indicators together with the measuring methods. These indicators are classified as applicable "regardless of the statistical production targets expected" or oriented to evaluate the quality in relation to specific statistical goals. At the international level attention should be drawn to the Statistical Network "Methodologies for an integrated use of administrative data in the statistical process - Administrative data (MIAD)", coordinated by Istat, which has deepened the different uses od administrative data and defined a framework for the quality of the administrative data, along with indicators for the scouting phase and for the acquisition phase and guidelines for their calculation. The ESSnet AdminData has produced a well thought-out list of quality indicators, for each Eurostat quality dimension, to be adopted when administrative data are used in the statistical outputs, with reference to business structural and short term statistics (ESSnet AdminData, 2013)[8].

The systematic calculation of indicators of the integration and treatment phases in the statistical process, specifically when these involve the use of administrative data, are less established. Many National Statistical Institutes, including the ISTAT, are working on the definition of indicators that reflect the errors described in the previous paragraph.

Some bibliographic references

Bakker B.F.M. (2010).Micro-integration: State of the Art. Note by Statistics Netherlands. UNECE Conference of European Statisticians. The Hague, The Netherlands, 10-11 May 2010

8 Tailored list of quality indicators: Structural Business Statistics (Annex 1a) e Tailored list of quality indicators: Short Term Statistics (Annex 1b).

Bryant J.R., Graham P.J. (2013). Bayesian Demographic Accounts: Subnational Population

Daas P., Ossen S. (2011). Report on methods preferred for the quality indicators of administrative data sources, Blue – ETS Project, Deliverable 4.2.

ESSnets Use of Administrative and Accounts Data in Business Statistics (2013). WP6 Quality Indicators When using Administrative Data in Statistical Outputs. Deliverable 6.3 / 2011: Guidance on the accuracy of mixed-sources statistics (available at https://ec.europa.eu/eurostat/cros/content/admindata-essnet-use-administrative-and-accounts-data-business-statistics_en)

Eurostat (2014). ESS Handbook for quality reports. 2014 Edition. http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

Eurostat (2011) "European Statistics Code of Practice – revised edition 2011". http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF

Eurostat (2009). ESS Handbook for quality reports. 2009 Edition.

Eurostat (2003a) "Definition of quality in statistics". Working group "Assessment of quality in statistics", Luxembourg, 2-3 October 2003. http://ec.europa.eu/eurostat/documents/64157/4373735/02-ESS-quality-definition.pdf

Eurostat (2002). Quality in the European Statistical System - The Way Forward, 2002 Edition (Leg on Quality) Luxembourg

Laitila T. Holmberg A. (2010). Comparison of Sample and Register Survey Estimators via MSE Decomposition (Q2014)

OECD OECD / Eurostat Guidelines on Revisions Policy and Analysis. http://www.oecd.org/std/oecdeurostatguidelinesonrevisionspolicyandanalysis.htm

ONS. Revision and correction policy (last update: July 2011). http://www.ons.gov.uk/ons/guide-method/revisions/revisions-and-corrections-policy/index.html

Sholtus S., Bakker F.M. (2013). Estimating the validity of administrative and survey variables through structural equation modeling. A simulation study on robustness. Discussion paper (201302). Statistics Netherlands.

Zhang L.C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica (2012) Vol 66, no.1, pp. 41-63.

Appendix. Guidelines for the centralised acquisition and management of an administrative data set

Principle 1. Discovery of new sources and their knowledge

It is opportune to have special attention to the emergence of new sources of administrative data, with respect to which it is essential to have an in-depth knowledge of all the legal and procedural aspects that regulate the life cycle of the administrative data, a precondition for their use within statistical production.

Guidelines

The European Statistics Code of Practice recommends reducing the statistical burden on respondents, and to improve the cost/ efficiency ratio in data collection (Eurostat, 2011). This supports the full exploitation of existing administrative data and alert to new sources not yet explored for use for statistical purposes.
The national legislation in question is in line with the above indications and requires the public authorities *"...that have archives, including computerised ones, containing data ... that are useful for statistical purpose, allow ISTAT access to these datasets and to the individual information contained therein… The access will occur according to the procedures as agreed to between the parties*. " (Law No. 322/89, as amended, Law No. 681/96, article 1, paragraph 8).

It is appropriate that there be, within the National Statistical Institute, an structured activity of "scouting" for new sources, functionally organised, that can also request the support of organisations put in charge of it, and which provides a sharing process with potential users within the Institute. An important contribution can be received by the communities of domain experts who can put their experience and knowledge at the service of the structures which manage the sources exploration.

To conduct this search of new sources, it is advisable to establish contacts with the various institutions operating on the area and to acquire information in a structured format[9] (D'Angiolini *et al*., 2014a; D'Angiolini *et al*., 2014a). This information should mainly cover the two hyper-dimensions pertaining the input quality: Source and Metadata (Daas *et al*., 2009, Iwig *et al*., 2013).

In particular, it is necessary to find information about the supplier institution, on the purposes and actual uses of the dataset, on privacy and security aspects, on conditions linked to the data supply, and on all the procedural and technical aspects relative to the acquisition, treatment and storage of data by the owner body, e.g., the used printed forms, the procedures for compiling the administrative act (whether compiled by the directly interested party or by an operator; whether it is on paper or computerised), the storage standards.

It is important to acquire all the conceptual metadata that allow the understanding of administrative data contained in the archive, i.e. a complete and accurate documentation of objects (units, events) and the variables contained in the archive, which may require a deepening of the legislation that regulates the life cycle of administrative data; for this reason it is important to make use of the support of the community of domain experts. Information about the period or date of reference of the administration data and the timeliness and frequency of archive updates are all particularly important.

[9] The ISTAT has started a co-ordination and harmonization of the printed forms activity, conducted through formal investigations about the administrative archives of the central institutions. The collected documentation is stored in a separate system called DARCAP (https://darcap.istat.it/darcap.php)

Some bibliographic references

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

D'Angiolini G., P. De Salvo, Passacantilli A. (2014a) ISTAT's new strategy and tools for enhancing statistical utilization of the available administrative databases. European Conference on Quality in Official Statistics (Q2014). Vienna, Austria, 2-5 June 2014

G. D'Angiolini, Patruno E., Saccoccio T., De Rosa C., Valente E. (2014b). DARCAP: A tool for documenting the information content and the quality of the available administrative data sources. European Conference on Quality in Official Statistics (Q2014). Vienna, Austria, 2-5 June 2014

Legislative Decree No. 322 of 6 September 1989, "Regulations on the National Statistical System and on the Reorganization of the National Statistical Institute, pursuant to art. 24 of Law No. 400 of 23 August 1988"

Eurostat - European Statistical System, Code of European Statistics, September 28, 2011. http://www.istat.it/it/files/2011/01/statistics_code.pdf

Iwig W., Berning M., Marck P., Prell M. (2013). Data Quality Assessment Tool for Administrative Data. http://www.bls.gov/osmr/datatool.pdf

Law No. 681 of 31 December 1996, "Financing of the Intermediate Census of Industry and Services in 1996"

Principle 2. Preliminary assessment on the suitability of the acquisition

The preliminary assessment on the suitability of the acquisition of an administrative dataset should be guided by an in-depth analysis of its current and potential relevance, of its cost/benefits ratio, of its stability and its expected quality.

Guidelines

The preliminary assessment on whether it is suitable or not to acquire an administrative dataset, or portion thereof, should be conducted on the basis of objective criteria, whether they are based on information from the investigations mentioned in Principle 1, if available, or *check-lists* made ad hoc.

The establishment of a round table, between the National Statistical Institute and the source owner body, is a support element in the decision-making process and the in the definition of possible agreements for acquisition procedures.

The evaluation should address several levels: the relevance of administrative data contained in the archive with respect to cognitive and/or production targets, acquisition costs, expected quality, cost/benefits ratio.

The analysis of the relevance of the administrative data, i.e. how it can be used for statistical purposes, should be performed by taking into account not only its current use, but also its potential use. Therefore, in this regard it will help both for the acquisition of information concerning the archive contents (legislation, reference collectives, used concepts, variables and their definitions and classifications), and the review of the possible uses among potential users, by the units in charge.

These uses do not regard, as is known, only the direct production of statistical information and the creation of statistical registers, but also the opportunity to improve the quality of certain production processes conducted within a National Statistical Institute. While, it might not be possible to completely replace a direct survey with data from administrative sources, it could still be convenient to acquire the archive for an indirect use, such as the creation/integration of *frames*.

For this purpose it is good to calibrate this preliminary evaluation, basing it on and differentiating it by type of use of the administrative data, especially when it is already known. In summary, the uses generally possible in a National Statistical Institute are: *i)* creation and maintenance of registers; *ii)* support to sampling designs; *iii)* replacement or supplementation during the data collection phase; *iv)* support to the editing and imputation procedures; *v)* support to the estimation process; *vi)*direct production of statistics; *vii)* support to the data validation (Statistics Canada, 2009).

The assessment about the acquisition of an administrative dataset can also derive exclusively from the opportunity to reduce the statistical burden on respondents.

The total costs of the acquisition and use in the production processes of the Institute (financial and in terms of instrumental and human resources used) can be difficult to predict. Among these are also to be included any costs related to the technological infrastructures, e.g. the development of exchange platforms. However, it is important to have an idea of the impact that the archive acquisition has on the Institute with respect to the above-mentioned costs. The calculated costs should be weighed against the achievable benefits not only in economic terms (reduction of direct collection costs), but also in terms of quality improvement (increased coverage, completeness of the lists, treatment of missing responses, both complete and partial responses, new statistics, more accurate estimates, validation, etc.).

Another factor which can affect the decision regarding archive acquisition concerns the stability in time of the administrative data contained in it, compared to the regulatory environment and procedures that regulate

its production. Frequent and significant changes in the structure, content, and archive data format could alter the cost/benefits ratio, with significant effects on the statistical production, the quality and time comparability of the data. This aspect takes on decisive importance for the statistics and registers produced with regularity and continuity.

Finally, the decision on acquiring the archive should also be based on an assessment of the expected quality, i.e. compared to the predetermined quality levels. It is worth emphasising that, at this stage, mention is made to the quality of the administrative dataset to be acquired, in the literature referred to as 'quality of the input ', and not quality of the statistical dataset derived after treatment process of the data contained in it, nor to the quality of the estimates inferred from subsequent processing. Furthermore, this assessment is independent from the statistical objective, which cannot be fully identified at this stage.

Therefore, at this stage, the evaluation of the expected quality of an archive to be acquired should be based on proven expectations in terms of:

- willingness from the institution to sign a formal agreement which establishes the procedures and times for possible data transmission, data security, documentation supporting the archive transmission;
- availability of comprehensive records of metadata (definitions for collectives and main variables);
- knowledge of technical conditions for the acquisition (accessibility, readability, conformity, data convertibility);
- knowledge of any archive deficiencies in terms of coverage/completeness in relation to the principal collectives and principal variables of interest for statistical production.

At this stage it is opportune to acquire a subset of trial data from the administrative dataset to verify its quality experimentally.

Some bibliographic references

Brackstone G.J. (1987). *Issues in the use of administrative archives for statistical purposes.* Survey Methodology, June 1987

Calzaroni M. (2011). *The administrative sources in processes and products of official statistics*, Istat http://www.istat.it/it/files/2011/02/Calzaroni.pdf (Last accessed: December 2013)

Daas P., Ossen S. (2011). *Report on methods preferred for the quality indicators of administrative data sources*Blue - ETS Project, Deliverable 4.2.

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

D'Angiolini G., De Salvo P., Passacantilli A. (2014).ISTAT's new strategy and tools for enhancing statistical utilisation of the available administrative databases. European Conference on Quality in Official Statistics (Q2014). Vienna, Austria, 2-5 June 2014 Statistics Canada (2009). *Statistics Canada Quality Guidelines*, Fifth Edition - October 2009

Tronti L. (2011) *The administrative data for statistics on local labor markets: the Guide project*. Public Function Department http://www.istat.it/it/files/2011/02/Tronti.pdf (Last accessed: December 2013)

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. Second edition: John Wiley & Sons, Chichester, UK.

Principle 3. Acquisition of an administrative dataset

The acquisition of an administrative dataset is regulated by agreements setting out the conditions regarding transmission, documentation and quality. There should be guaranteed a close monitoring of the legislative or procedural changes having an impact on the structure or data of the archive. Those amendments should be promptly communicated to the internal users.

Guidelines

The administrative dataset should be acquired through the establishment of formalised agreements with the owner body (conventions, memorandum of understanding, etc.). These agreements should establish: the methods and times of data transmission, the documentation supporting the transmission and contents of the archive, the rules for the respect of confidentiality and also the procedures for returning the statistical information to the supplying institution.

In detail, as suggested by the UNECE (2011), it is advisable that these agreements:

- contain a reference to the legislation, if any, which allows the Statistical Institute to access to the data;
- precisely identify the people and facilities for transferring and receiving of the archive;
- identify all the metadata and information about the quality that must represent the basic documentation for the correct use of the dataset. In particular, important information pertains to: the definitions of the dataset objects (units, events) and the variables and classifications used, time references of the dataset data, descriptions of, if any, treatments that the data has undergone before being transmitted to the Institute;
- require a detailed description of the main populations and variables derivable from the dataset and quality in terms of coverage of the populations and completeness of the variables;
- lay down the timing (first delivery date, deadline of subsequent supplies) for transmission;
- lay down the rules and procedures pertaining to the protection of confidentiality which ensure the transmission methods and the treatment of sensitive data, and the prevention against the risk of confidentiality breach;
- set the validity terms of the agreement;
- establish supply costs, if any;
- determine the conditions for the possible periodic supply of the dataset;
- require the supplying institution to give timely notice of any changes in the structure and/or data content as a result of legislative changes or other reasons;
- require the Institute to a return of information in the form of statistical data as a form of interchange, as part of the collaboration built in time, between the reference administration and that of the Institute;
- provide a technical annex, so that the administrative dataset is transmitted securely and through protocols conforming to the Institute's standards.

The identification of a contact person at the owner body and the return of information to the supplying institution are elements that set the basis for more use of administrative data. Improving the climate of cooperation between the institutes involved in the acquisition process could make it possible to have an active role in the design phase of the printed forms used for data collection, among other things to limit the consequences of any changes in the reference regulations and, therefore, to continue to ensure over time the comparability of the statistics produced by the Institute.

It is important to ensure timely communication by the owner bodies of the administrative data regarding changes made to the printed forms used and in the definitions of administrative concepts (D'Angiolini *et al.*

2014). The support of the domain experts community can provide a better understanding of the impact of such changes on the statistical production.

Some bibliographic references

Australian Bureau of Statistics (2011). Quality Management of Statistical Output Produced from Administrative Data. Information Paper. Australia. March 2011.

D'Angiolini G., Patruno E., Saccoccio T., De Rosa C., Valente E.(2014). DARCAP: A tool for documenting the information content and the quality of the available administrative data sources. European Conference on Quality in Official Statistics (Q2014). Vienna, Austria, 2-5 June 2014

Statistics Canada (2009). *Statistics Canada Quality Guidelines*, Fifth Edition - October 2009

Statistics Canada (2009). *Statistics Canada, Use of administrative data (website)*http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm (Last accessed: December 2013)

United Nations Economic Commission for Europe - UNECE (2011). *Using administrative and secondary source for official statistics: a handbook of principles and practice*. United Nations, New York and Geneva, 2011.

Principle 4. Pre-treatment, quality controls and release of the administrative dataset for further uses

During the administrative dataset acquisition, the documentation and quality are verified and possibly integrated with information that can increase the clarity and usability of the archive. The pre-treated administrative dataset should be made available for internal use, complete with all the metadata and quality indicators on data before and after treatment.

Guidelines

The archive is acquired by the Institute according to the procedures stipulated with the owner body through the department in charge of: handling relations with suppliers, carrying out general quality controls, performing pre-treatment on data and preparing the datasets for subsequent integrations.

In general, the pre-treatment of the received data can be subjected to the following phases: *i)* conceptual analysis of the objects contained in the dataset; *ii)* uploading of data in internal data-bases; *iii)* identification of units and assignment of univocal identification codes that are stable over time (univocal keys); *iv)* standardisation of a subset of data; *v)* recoding of a subset of variables. Finally, quality indicators are calculated and documentation on the metadata and quality is prepared (Di Bella G. and Ambroselli S., 2014; Kobus P. and Murawski P; Cetkovic P., *et al*., 2012).

The acquired dataset should be included in the Institute's current production system; this implies that the format in which the dataset is provided should be made compatible, if possible, with the formats in use at the Institute. In the acquisition phase, the related controls should be performed: the correspondence between the number of imported records and those expected (also based on previous supplies); the clarity of the record layout; correspondence between the record layout and imported data (to avoid possible misalignment of the columns); the existence of univocal keys that are comparable with those in use at the Institute; the correspondence between the type of variable (numeric, alphanumeric) and data format; the adequacy of the length of the fields assigned to the variables. For specific measures, in this phase, reference can be made to the indicators present in the size of the *Technical checks* defined in the BLUE-Ets project (Daas *et al*., 2011).

Then, record duplicates, if any, have to be removed. This may require that the identification variables of the record are previously subjected to standardisation procedures ("*parsing*", i.e. separation of a certain variable into several variables, as is done to standardise addresses or the names and last names of individuals).

Therefore an analysis should be conducted of the classifications used in order to understand their degree of compliance with classifications that are standard or in use at the Institute. In case of non-correspondence between classifications, it is necessary to integrate data with relevant classifications.

It is worth noting that, at this stage, the controls applied are aimed to obtain elements to decide whether or not the provided dataset can be considered valid and releasable to internal users in complete or partial form or with notes of caution regarding certain variables. In general, at this stage, editing and imputation procedures typical of data treatment for statistical purposes are not applied.

In case of evidence of non-compliance of the supply, it is desirable to check with the owner body of the administrative dataset.

It is advisable to attach a quality report to the dataset that includes a description of the metadata and the transformation and integration process undergone, and some of the general quality indicators (coverage of the principal reference populations, rates on missing data for principal variables, etc.).

The information regarding the available version of the dataset, the documentation on the pre-treatment procedures and quality controls performed should be made available to support internal use of the dataset.

After identifying the procedures for validation of the administrative data being received, it is advisable that they are applied regularly on subsequent supplies.

The descriptive metadata of the administrative dataset acquired should be consistent with those defined by the Unitary System of Metadata (SUM). For the quality reporting, reference can be made to the Quality Report Card, which contains indicators extensively tested and validated internationally (Cerroni *et al*., 2014).

Some bibliographic references

Cetkovic P., Humer S., Lenk M., Moser M., Schnetzer M., Schwerer E. (2012).*A quality monitoring system for statistics based on administrative data.*European conference on Quality in Official Statistics Q2012, 29 May - 1 June 2012, Athens, Greece.

Cerroni F., Di Bella G., Galiè L. (2014). *Evaluating administrative data quality as input of the statistical production process.*Rivista di Statistica Ufficiale, issue no. 1-2, 2014

Daas P., Ossen S. (2011). *Report on methods preferred for the quality indicators of administrative data sources*Blue - ETS Project, Deliverable 4.2.

Di Bella G., Ambroselli S. (2014). Towards a more efficient system of administrative data management and quality evaluation to support statistics production in ISTAT. European Conference on Quality in Official Statistics. Q2014. Vienna, Austria 2-5 June 20014

Kobus P. and Murawski P. (not known). Transforming administrative data to statistical data using ETL tools. http://www.ine.es/e/essnetdi_ws2011/ppts/Murawski_Kobus.pdf

Statistics Canada (2009). *Statistics Canada Quality Guidelines*, Fifth Edition - October 2009

Principle 5. Monitoring and evaluation of the dataset and feedback to the supplying institution
The use of the acquired administrative dataset should be monitored and evaluated through a coordinated and shared process, in order to identify both shortcomings in terms of quality and treatments common to several processes. The results of monitoring and evaluation should be shared with all current and potential users and possibly transmitted, with appropriate procedures, to the archive owner body.

Guidelines

It may occur that the same administrative dataset is used as input in several production processes. In particular there may be cases in which sets of variables or different subpopulations are used in several production processes, or that the same variables constitute the input for the production of several statistics (e.g. short terms and structural statistics). The type of use can vary when administrative data are used directly or in support of the survey process.

Therefore, it is necessary to carry out periodic checks on the types of use of administrative data in the Institute namely, for each production process, to know which administrative data are used and how, preferably through the exchange of information between the various informational systems of process management, with a view to inter-operability that optimizes the overall efficiency. In this way you can get a complete picture of the relevance of each source on the basis of which to define the particular levels of attention necessary to manage the "dependence" of the institute's production from administrative data.

From a management point of view it is also advisable to make use of evaluations by the community of experts and users of administrative sources, especially for the sources most utilised by the Institute, such as the social security and tax data. The coordination has to be managed centrally in order to share the common problems, promote synergies and avoid duplications of activities with a view to efficiency and standardisation.

Finally, the sharing of possible data quality problems, that may limit their statistical use, allows agreements to be made with the supplying institution regarding actions aimed at overcoming them. From the simplest case of increase of data supply timeliness, to the case of improvement in the clarity of the metadata or the possible reduction of the distance between administrative concepts and statistical concepts. This feedback process is particularly important in order to enhance the use of administrative data in time.

On the other hand, the coordination of internal users of the source makes possible to precisely define the set of data required in relation to the periodicity and timeliness of the various supplies and avoid unnecessarily burdening the activities of the source holder for the supply of data.

Some bibliographic references

Di Bella G., Ambroselli S. (2014). Towards a more efficient system of administrative data management and quality evaluation to support statistics production in ISTAT. European Conference on Quality in Official Statistics. Q2014. Vienna, Austria 2-5 June 20014.