**Quality Guidelines for statistical processes using administrative data**

Barbalace F., Boggia A., Brancato G., Busetti C.

Italian National Institute of Statistics Istat

As many Statistical Institutions, Istat is striving for a great exploitation of administrative sources for statistical purposes. As a consequence, the use of administrative sources in different steps of the statistical business processes and for different purposes is increasing. Such an investment requires the development of tailored tools for quality assessment. Istat quality assessment strategy currently includes, among the others, statistical auditing and self-assessment procedures, aimed at evaluating compliance towards standards, i.e. Quality Guidelines for statistical processes, publicly available on Istat website. These guidelines are suitable for carrying out evaluations only on a subset of processes using administrative sources and require to be extended to cover other processes using administrative sources. The paper will describe quality issues in administrative data and how the quality guidelines are being developed in order to allow a wider and more focused application of statistical auditing and self-assessment procedures within Istat. An analysis of Istat processes using administrative sources will also be presented. The followed approach is taking into account for Istat organizational structure as well as for the different activities that are being carrying out on this relevant theme across the Institute.

1. **Introduction**

The needs for new and updated statistical information is increasing every day as well as the availability of administrative sources ruling various kinds of fields. At the same time the increasing use of new sources of administrative data is also supported by the opportunity to reduce the statistical burden on respondents and to improve the ratio of cost/effectiveness in data collection, as recommended in the European Statistical Code of Practice (ESS CoP).

Consequently, Istat has greatly increased the use of administrative data sources that are annually acquired, that have doubled in the last five years and that has reached about 230 data sets nowadays. This paper presents the work of development of Quality Guidelines for statistical processes using administrative data, that are thought to integrate the already available Istat quality guidelines for direct surveys, produced in 2011 and currently used as compliance standards in the auditing and self-assessment procedures.

Section 2 illustrates the underlying quality model around which the guidelines have been developed and specifies the fields covered. In Section 3 the main issues included in the guidelines are summarized. Section 4 reports the current available documentation at Istat with respect to the statistical production using administrative sources. Finally, conclusions are drawn in Section 5.

The quality guidelines for processes using administrative data are being developed based on a wide range of bibliographic references, specific for each issue that is being faced, and that is not feasible to report in this paper. Therefore only the very few papers representing pillars of the work and of this paper are listed at the end.

1. **The Quality model for statistics using administrative data**

In order to structure the Guidelines, a model for quality has been identified. This is based on the definition of the following main areas:

* Usability of administrative sources
* Input quality
* Throughput quality
* Output quality

The usability refers to the set of overall assessment aimed at evaluating the general possibility to use the administrative sources for statistical purposes. It cannot be strictly considered a quality evaluation, since the administrative register is not originally created to draw statistics. However Istat is carrying out many activities in this field aimed at coordinating the administrative data production and increasing the quality culture among producers, useful to enlarge the number of potentially usable sources and their quality. This area is not covered by the Quality Guidelines. The quality Guidelines are indeed developed around the three main areas: input, throughput, output. The first refers to the quality assurance of data that are centrally acquired, controlled, pretreated and made available for the specific use within the statistical production processes (throughput quality). Output quality refers to the quality of the final statistics derived using administrative data.

Considering input quality, it is to be noticed that, respect to the common use of the term “input quality” [3], Istat guidelines consider also issues related to the internal management and monitoring.

With respect to the throughput quality, starting from the approach of Zhang [7] and the main treatment steps listed in Wallgren & Wallgren [5], the errors that can be generated during the main phases and activities have been identified and principles and guidelines developed accordingly. They are summarized in Figure 1. The activities performed depend on the type of use of the administrative data and the order is not strictly defined. Not all the phases are performed in each process.

The potential errors can be generically referred to the administrative data objects that assume the role of units (e.g. individuals, businesses) or events (births, marriages, ...).

**Table 1. Throughput quality: Main phases/activities of the statistical production processes using administrative data and potential errors**

|  |  |
| --- | --- |
| **Main phases and activities of the processes** | **Potential errors** |
|   | **Objects (units, events)** | **Variables** |
| Identification of the target population and concepts  | Coverage error  | Validity / specification error  |
| Identification of available sources | Comparability error |
|   |  Missing information |
| Use of administrative data as population | Coverage error: missing, duplicates, delays |
|   |
| Integration | Linkage error: false links, false non-inks | Measurement error: strict measurement errors, mapping errors, compatibility errors, comparability errors |
|   |
| Units harmonization  |  |
| Variables and classifications transformations or derivations |
| Time dimension alignment |
| Editing and imputation & Micro integration |
| Estimation |

The first three phases/activities of Table 1 are related to the evaluation of the suitability of the use of administrative data form a conceptual, comparative and operational perspective. If the target statistical concepts do not match with the administrative ones, coverage errors may arise (referred to the units) and validity or specification error can be generated (referred to the variables). The best available source to be used is then identified and geographical heterogeneity or over time instability of the legislation may affect the comparability of the data (comparability error). Using the administrative register as the whole or a part of population can give rise to missing units (or events) and duplicates and missing information. Also delays in registering events or in cancelling them have an impact on undercoverage or overcovage, respectively.

Integration may lead to integration errors, i.e. false links and false non-links. The former is due to the linkage of records referred to different units as if they were the same unit. As a result on the observed variables, measurement and compatibility errors might rise [7]. The latter, in general, may result on overcoverage on the units or, if integration is aimed to impute variables, also to missing values.

Harmonizing units, variables and classifications can be straightforward or require complex transformations. The potential errors have been included in a wide class applied to the variables, referred to as measurement errors, that includes: strictly measurement errors (errors in the data collection from the administrative body), mapping errors (errors in re-classifying measures), compatibility errors (derived from inconsistencies among sources) and comparability errors, as already defined. Again, unit harmonization may entail coverage errors. It is important to underline that coverage errors may also derive from operations of integration and harmonization, as attempted to be shown by the arrows included in Figure 1. Finally, although Editing and Imputation is a way to remove measurement errors, it can itself introduce some.

In the schema, the usual phases of validation of the results, archiving and documenting are not included, since their impact on quality is expected to be limited. They are included in the guidelines.

Considering output quality, being the output represented by statistics, the widely shared quality components, defined at European Statistical System, are considered taking into account for the impact of using administrative data.

1. **The content of the guidelines**

**3.1. Input quality and centralized management**

This area of the guidelines concerns different steps for the acquisition of an administrative register: from the discovery of new administrative data sources, to the decision on their acquisition, to preliminary quality controls and pre-treatment of the data in order to provide the archive to internal users, to the monitoring on the internal uses and users.

Emerging statistical needs impose an activity of “scouting” for new data sources, carried out with the support of internal experts and potential users, and based on information from any investigations already available, or on the administering of a tailored “check-list” addressed to the owners [4]. Every new data source requires a thorough knowledge of all elements of legislation ruling the whole life cycle of the administrative data, that is a precondition for the correct use of the administrative data in statistical production processes.

The guidelines recommend an activity of continuous monitoring of the administrative context in which the data are generated. It is also recommended to acquire a thorough understanding of the conceptual elements underlying the register (units, events, variables and reference period) as well as the procedures (forms and collection technique for the administrative data, frequency of register update).

The evaluation on the acquisition of the administrative register should be based on the information available on objective criteria. This implies to take into account current and potential relevance of the data contained in the database, such as potential uses for direct statistical tabulations or support to other statistical processes, and last but not least the trade-off between costs and benefits. The expected quality of the data should also play an important role in the decision on the acquisition. The concept of expected quality should be intended in a wide sense, included long-term stability of administrative data supplied: in fact, frequent and significant changes in legislation, structure, content, and format of data in the register could alter the cost/benefit ratio, with significant impact on the quality and comparability of data over time. However, it should be noted that this step is referred to input quality dimension of the administrative register for statistical purposes and not to its transformation into statistical register. To evaluate input quality, indicators proposed in the literature can be used, as some of those reported in BLUE-Ets [3].

When the administrative register is acquired, good relationships with the agencies providing the administrative archive, should be established and maintained, and formal agreements should be set up. They should: cover procedures and timing for data transmission, set data quality standards, establish the required documentation, and commit the institute to provide back the derived statistical information. Then, the compliance to the agreed standards and the availability of proper documentation should be evaluated and some technical checks should be applied to the data received.

The register is then integrated with information and transformations which increase its own clarity, readability and usability, and made at disposal of the internal structures of the Institute, supported with all available metadata and quality indicators, under the form of a Quality Report Card.

Finally, for each acquired administrative register, it is recommended a continuous monitoring on internal uses and users. Regular investigations on internal users satisfaction regarding usability and data quality during the integration and use of the register in specific statistical processes, allow an objective evaluation of the register quality and permit to fruitfully provide feedbacks to the suppliers, with the aim of improving the quality of the input.

To be noticed that the general principles stated in this section of the Guidelines are being carried out within Istat by a centralized unit, in charge of the acquisition, pre-treatment, quality checks and monitoring of administrative sources (ADA, Administrative Data Acquisition and integration).

**3.2. Throughput quality**

Regarding throughput quality, the focus shifts from the central acquisition to the uses of the administrative data within the statistical productive process for specific estimation objects. In particular, principles and guidelines are developed following the phases and activities drawn in Table 1.

*Identification of administrative sources and their use*. Within a statistical production process, administrative data can be used in different ways and for different purposes. They have been widely described in the literature and are not repeated here. Since in general the administrative data will never be used as they are, but they will require some transformations, as initial step, the guidelines recommend focusing attention on what type of use is foreseen, carrying out a thorough analysis of the process steps involving the use of the administrative data, because this orients the methodologies for data treatment and the quality evaluation methods. The compliance of administrative concepts against the statistical ones, coverage issues, data quality, conditions on the register acquisition, are all elements to be evaluated in the light of the specific statistical use.

The choice of the most suitable administrative source to use, if applicable, can be guided on the basis of comparison studies, taking into account not only the advantages of the use of the administrative data but also the risks deriving from possible losses of information due to instability of the source.

*Methods for data integration.* Within the statistical production processes that uses administrative data, the integration between data sources may have different purposes and characteristics. Simply put, the integration can be used: to build micro-data archive for the direct production of statistical data or for the support of other productive processes (as frame, for data validation, etc. ) or to complete information on populations and/or variables.

The guidelines focus attention on the record linkage (RL) procedure, both for integration between administrative sources, and also, between administrative sources and survey data, orienting on the steps to be followed for the application of the procedure.

RL between administrative sources can be done in different ways, principally deterministic or probabilistic. The next step relates to the selection of the matching variables able to jointly identify the units of the population of interest. When administrative sources have a unique error-free key, then it is possible to consider exact matching techniques. When the unique key is not present, generally, key variables are selected, which when jointly considered may contribute to identify the units (e.g. name, date of birth, address…). The presence of possible errors in the identification keys increases the possibility of generating errors during this phase, which is extremely important in the use of administrative data. A further step may be related to the choice of eventual blocking and file sorting operations aimed at reducing the computational problems due to the size of the possible comparisons. Couples are assigned to matches or non matches, based on a comparison function.

Finally, the guidelines recommend the evaluation of the quality of the whole process of integration, calculating appropriate indicators. Whenever possible, in RL applications the false link rate (units erroneously matched by the procedure) and the false non link rate (units erroneously not matched by the procedure) should be estimated, also recurring to manual checking, performed on a sample basis.

*Identification and transformation of units, variables and classifications.* Two principles and the relative guidelines are devoted to the activity of harmonization of statistical units, variables and classifications from administrative objects into statistical concepts.

The process of identification and mapping of the administrative objects to the statistical ones may require alignment procedures with a different degree of complexity, form a simple situation in which there is a direct one-to-one relationship, to situations in which integration among registers and manual support of experts are required. Further complexity is introduced when for the target statistical units also the relationships with other units have to be considered, as for example in many cases for business statistics.

Once the unit alignment is performed, the coverage error should be estimated and an attempt to understand the causes of the error, according to the model presented in Section 2, should be made (conceptual, registration lacks or delays, integration errors). The methods to assess coverage are known from the literature, e.g. capture-recapture, but sometimes they require many assumptions and are difficult and expensive to apply.

Concerning the variables and classifications, similar problems need to be faced. Methods are based on transformation functions that may allow exact calculations, or may assume deterministic rules or may be more complex, being based on casual models or combinations of variables included in different registers. Again, whichever is the method, an assessment of the potential errors due to this activities should be pursued. Methods and capacity to obtain error estimations depend highly on the available information, and range from comparison studies with survey or census data to more complex statistical models, e.g. structural equation modelling, when no correct available information (*gold standard*) is at hand.

Finally the impact on final results of using data from administrative sources as compared to survey or census data should be assessed.

*Time dimension.* The different characteristics related to time of the administrative register have to be considered, and the impact on the statistical result evaluated. First of all, the reference time of the administrative data, that may be different from that related to the target statistical concept, thus requiring transformations. Then the time lag between the event and its registration in the register, that may play an important role in deciding if using the register within the statistical process and also may guide in the choice of the opportune timing to integrate the register in the production process. Also the nature of the administrative data, if flows or stocks, may influence the evaluation.

*Editing and imputation.* A relevant phase of the statistical process is the editing and imputation (E&I) procedure, which has to be properly tailored when the administrative data are used and potential errors due to the use may arise as a consequence. Indeed, if in general, the main objectives of a procedure of E&I, which are applicable to a statistical production process, are also valid in the context of the use of administrative data, it is necessary to consider the peculiarities of these data, because they have an impact on the organization of an E&I strategy.

The first element to consider is whether the statistical data are obtained by using a single source or derived from the integration of more sources. Concerning the use of a single source, the design of E&I procedure is similar to that of a typical statistical survey. In the case of integrating more sources, the strategy of E&I procedure is more complex and offers different alternatives: *i)* E&I performed first in each source and then in the integrated data; *ii)* E&I applied directly in the integrated data. The two approaches have to be evaluated on the basis of expected quality results in the data and the resources necessary for the procedure.

Moreover, the guidelines focus attention on the specific error detection and correction methods, highlighting limits and possibilities deriving from the administrative nature of the data in the application of the known methods for editing and imputing.

Finally it is recommended to evaluate the quality of the E&I procedure by calculating appropriate indicators on the row and cleaned data.

*Archiving, Dissemination and Documentation.* Finally, archiving of data and metadata should follow the standards adopted by the Institute and proper documentation for internal and external users produced. In order to assure transparency, it is advisable that such a documentation clearly states which administrative data have been used for statistical purposes and how. Internal users may be interested in all the quality indicators, including input quality measures, external users will in general be more interested in measures of the quality of the output

**3.3. Output quality**

By output or product quality it is intended the quality of the statistical information produced by the statistical production process that uses administrative data. It does not include outputs represented by intermediate data or microdata sets used for direct tabulations. This section is focused on the use of administrative data for substitutions of units and/or variables, excluding therefore the situations in which administrative data are used as frame or for data validation.

Being statistical in nature, output quality is evaluated and measured according to the commonly used and accepted quality dimensions (Relevance, Accuracy, Timeliness and Punctuality, Coherence and comparability, Accessibility and Clarity). Nevertheless, the use of administrative sources have an impact on: *i)* the applicability and our capacity to derive measures for the quality dimensions; *ii)* the levels of output quality.

With respect to the first aspect, some dimensions are measured exactly as they are for direct surveys (timeliness and punctuality, accessibility and clarity). Accuracy is, by all means, the more demanding to consider and depends on the estimators that are applied to combine administrative and survey data. In this respect the guidelines make reference to the applications available from the literature and to the more recent studies and attempts to derive accuracy measures.

With respect to the second point, indeed relevance, coherence and comparability are the dimensions more affected by the use of administrative data. Relevance is strongly dependent on possible validity and measurement errors generated in the process of transformation of administrative concepts. Coherence and comparability may be highly affected by the stability of the legislation ruling the administrative data.

1. **Istat statistical production using administrative data**

As direct statistical processes, also processes using administrative data are currently documented in Istat Information System for quality and metadata documentation (SIDI/SIQual) [2], according to the common uses. The system documents, for different kinds of processes (direct, using administrative data, partly direct partly using administrative data or “mixed” surveys, statistical compilations), metadata on process content, survey phases and operations, and activities developed to guarantee the quality of statistical information in terms of prevention, monitoring and evaluation of sampling and non-sampling errors.

Considering the use of administrative data to totally replace the population, SIDI/SIQual documentation system distinguishes between the situation in which a single body is the owner of a central register from the situation in which several institutions act as reporting units.

A description of the metadata documented for processing having at least an administrative component, in addition to the process metadata that are in common with the direct surveys, is reported in Table 2. They reflect the more significative characteristics concerning the administrative register as identified also in Brackstone [1].

Regarding processes using administrative data for other purposes, such as to improve the quality of survey data once it is collected, check for errors, impute missing data and adjust the estimates, this kind of information can also be documented in the system.

**Table 2. SIQual metadata documentation on origin, acquisition and quality control of input sources**

|  |  |
| --- | --- |
| **Documentation aspects** | **Information elements included** |
| ***Origin of the administrative data*** | ***Act ruling the administrative record****: law reference, periodicity of the administrative event, geographical homogeneity of the regulation, satisfaction on the administrative data with regard to some quality dimensions.****Record formation****: paper or electronic record; person in charge of the compilation; type of requirement: legal requirement to register an event (e.g. medical file); benefit or obligations administration (e.g. balance sheets); administrative variations (e.g.: new municipalities); instrumental measurements (e.g. air pollution).* |
| ***Data acquisition*** | ***Type of data****: individuals vs. aggregated records, pure administrative vs. statistical data collected while drafting administrative records.* ***Data provider***: *a central institution providing a register, many institutions as reporting units, intermediate bodies.* |
| ***Data transmission*** | ***Mode and means****: paper or electronic transmission (e.g. on electronic devices, File Transfer Protocol on public network, via Istat or others websites); generalized software used for data capturing.* |
| ***Quality control actions*** | ***Control by the providers:*** *control on data completeness and accuracy .****Control by Istat:*** *existence of an agreement between Istat and the body in charge of administrative data; on-line courses for intermediate bodies/institutions in charge of data collection; control on nonrespondent bodies/institutions; control on update and coverage level of body/institution frame; quality control on information on the surveyed analysis units.* |

Analysis of the distributions of surveys using administrative data and mixed surveys according to the metadata documented are possible. Currently 152 primary surveys, 46 out of which are using exclusively administrative data and 9 mixed, are carried out by Istat, meaning that more than one third of primary surveys are based on administrative sources. As expected, the use is even higher when considering the statistical compilations.

Table 3 reports the distribution of the surveys using administrative data according to the typology of provider.

**Table 3. Istat primary surveys acquiring administrative data by type of survey and provider**

|  |  |
| --- | --- |
| **Data provider entity** | **Type of survey** |
| ***Admin. Sources*** | ***Mixed*** | ***Total*** |
| *A central institution providing a register (A)* | *4* | *2* | ***6*** |
| *Many institutions as reporting units (B)* | *36* | *4* | ***40*** |
| *Both A and B* | *6* | *3* | ***9*** |
| ***Total*** | ***46*** | ***9*** | ***55*** |

The system also allows to monitor the data transmission, showing that there is a transition process towards the electronic means, testified by the decrease of the number of surveys acquiring data exclusively on paper that passed from 16 in 2011 to 10 today (data not shown).

Regarding data acquisition, in 24 processes the administrative form is defined by Istat or shared with Istat, thus increasing the capacity to collect the target statistical concept.

The SIDI documentation system also stores standard quality indicators [2]. Few of them are output oriented, thus referred to the statistical product (timeliness and punctuality, comparability, revisions, coherence between provisional and final results, coherence with other sources), others are process oriented (coverage, unit nonresponse, coding, editing and imputation). They all are applicable in processes that use administrative data. Some of them may not be interesting, as for example, in some cases, response rates referred to reporting units that are administrative bodies. In general, from the analyses carried out on quality indicators available from the SIDI system, processes using administrative sources are scarcely affected by unit nonresponse errors while are negatively influenced by timeliness problems. For many surveys using administrative data coherence indicators are computed, comparing results with those from other sources. It is the case of some demographic phenomena for which annual and monthly quick surveys are carried out on the same statistic. Finally, it is not the aim of SIDI to have specific quality indicators on input quality [3], that are not implemented, but are being computed by Istat sector in charge of the acquisition of administrative registers.

1. **Conclusions and future work**

In this paper the scopes and the general content of the Quality Guidelines for statistical processes using administrative data have been illustrated. The work is almost completed although different sections of the guidelines have different degree of maturity and stability. Indeed, it is still under evaluation the opportunity to have a principle and related guidelines on the estimation phase.

This work represents the precondition for future activity. Once a first version of the Quality Guidelines is released, the tools supporting the statistical auditing and self-assessment will be developed. They are questionnaires (to be administered and for self-administration) able to assess the compliance towards the principles stated in the guidelines and the template for an evaluation report with improvement actions to be filled in after the assessment exercise.

Istat auditing and self-assessment procedures are extensively based on the quality documentation stored in SIDI/SIQual. Therefore, in order to generalize the assessment to processes using administrative sources, also a critical analysis of the suitability of the documentation structure and content in the system will be required.

**Acknowledgements**

The drafting of the Quality Guidelines is coordinated by G. Brancato. Up to now, the guidelines have been written by: Barbalace F., Boggia A., Brancato G., Cerroni F., Di Zio M., Filipponi D., Luzi O., Scanu M.

The Guidelines are subjected to an internal review process. Up to now the guidelines have been revised by: Di Bella G., Pallara A., Signore M.

**References**

 [1] Brackstone G.J. (1987), Issues in the Use of Administrative Records for Statistical Purposes, Survey Methodology, 13, 1, 29--43.

[2] Brancato G, Pellegrini C, Simeoni G. e Signore M. (2004), Standardising, Evaluating and Documenting quality: the implementation of Istat information system for survey documentation – SIDI”, Proceedings of the European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz, 24-26 May 2004.

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

[4] 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.

[5] Wallgren A., Wallgren B. (2007), Register-based Statistics. Administrative Data for Statistical Purposes, John Wiley & Sons, West Sussex, England.

[6] Zabala, F., Reid, G., Gudgeon, J. and Feyen, M. 2013. Quality Measures for Statistical Outputs using Administrative Data. Statistical Methods, Statistics New Zealand.

[7] Zhang Li-Chung (2012), Topics of statistical theory for register-based statistics and data integration, Statistica Neerlandica, Vol 66, Nr. 1, 41--63.