This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Editing Administrative Data

Contents

General section

1. Summary

The use of administrative data as a source for producing statistical information is becoming more and more important in Official Statistics. Several methodological aspects are still to be investigated. This module focuses on the editing and imputation phase of a statistical production process based on administrative data. The paper analyses how much the differences between survey and administrative data affect concepts and methods of traditional editing and imputation (E&I), a phase of the production of statistics that nowadays has reached a high level of maturity in the context of survey data. This analysis enables the researcher to better understand how and to which extent traditional E&I procedures can be used, and how to design the E&I phase when statistics are mainly based on administrative data.

2. General description

The use of external information in statistical production processes is increasing its importance in the National Statistical Institutes (NSIs).

External information generally refers to *secondary data*, i.e., data not collected directly by the user. An interesting discussion on the use of this kind of data can be found in Nordbotten (2012). In this paper, the focus is on administrative data, which is a subset of secondary data. They have the characteristic of being collected for non-statistical purposes and at the moment they are the mostly used external source of information in NSIs.

Administrative data are collected for administrative purposes, e.g., to administer, regulate or tax activities of businesses or individuals. Although not yet fully explored from a methodological point of view, the field of the statistical use of administrative data can be considered in an advanced state for a number of critical issues like accessibility, confidentiality and risk of misuse.

The usefulness of administrative data depends on their concepts, definitions and coverage (and the extent to which these factors stay constant), the quality with which the data are reported and processed, and the timeliness of their availability. These factors can vary widely depending on the administrative source and the type of information (Statistics Canada, 2010).

It is worthwhile to remark that, although this definition could be applied to survey data, in the context of administrative data it assumes a particular importance since most of the elements considered in the statement are not under the control of the NSIs, while on the contrary for survey data NSIs can, at least in principle, design opportunely all or most of them.

The main advantages deriving from the statistical use of administrative data include: the reduction of costs (in the long term) and of respondent burden, deriving from the reduction of information needs from direct surveys; the improvement of timeliness and accuracy of statistical outputs; the increased potentials for more detailed spatial-demographic and longitudinal analysis.

Main drawbacks are connected to the initial costs due to gain access to the new sources, matching classifications, harmonising concepts and definitions with respect to the target units and the statistics of interests, and assessing quality. Concerning the latter aspect, it is worthwhile noting that the quality of data collection, data capture, coding and data validation are under the control of the administrative program and may focus on aspects that could be not relevant for the NSI's purposes. In general, these validation activities cannot be considered sufficient to ensure the statistical usability of the data, and extensive additional data editing activities need to be performed before incorporating external data into statistical processes. Methods and tools are to be developed to this aim taking into account the peculiarities of administrative data. In addition, the use of an administrative source generally implies the need of other sources (including surveys) to compensate for non-covered units/variables, thus editing strategies for multi-source data should be developed.

The impact of using administrative data in statistical production processes depends also on their supposed use. Two different scenarios can be distinguished:

1) administrative data support surveys: they are used to maintain frames, to improve the efficiency of sample surveys (calibration), to provide information which might be used to assist the E&I process, as an information source that might be used for quality assurance (for instance to compare results);

2) administrative data serve as a source for providing the statistical output required, in this case they can be used as a primary source or by integrating them with survey data.

In this paper the focus is on the use of administrative data under scenario 2.

The paper is structured as follows. In Section 2.1 the main objectives of data editing for survey data are discussed in the framework of administrative data. Section 2.2 is dedicated to the illustration of error characteristics in administrative data. The application of traditional methods used E&I is discussed in Section 2.3. How to provide information about data quality is illustrated in Section 2.4. General ideas about the design of E&I of administrative data are proposed in Section 3.

2.1 Statistical data editing of administrative data

The main objectives of statistical data editing are reported in the following list (cf. "Statistical Data Editing – Main Module"):

OB1 To identify possible sources of errors so that the statistical process can be improved in the future;

OB2 To provide information about the quality of the data collected and published;

OB3 To detect and correct influential errors in the collected data;

OB4 To provide complete and consistent data.

When discussing E&I for administrative data, the main question is how much the concepts developed so far for E&I of a single statistical survey (see EDIMBUS, 2007) can be translated into the administrative data framework. The question is translated in two main questions: 1) whether the above mentioned objectives are still valid, and 2) whether error characteristics and methods usually adopted for detection and treatment are the same. To give an answer to those questions, differences between administrative and survey data should be highlighted.

Two important distinctive characteristics are:

i. the process of gathering information is not generally under the control of the entity (for instance the NSI) that will provide the final figures,

ii. information is gathered for other purposes.

Other important differences are that:

iii. generally the sizes of the data bases concerning administrative data are much larger than those concerning survey data,

iv. administrative data are frequently used in a statistical production process where data sources are combined and integrated. The integration of data sources becomes a specific trait of the use of administrative data since, as they are gathered for other purposes, they generally do not observe all the variables of interest, and most of the times they refer to a population covering a part of the target population. In those cases, integration between administrative sources and surveys is required to fill the gaps.

Those peculiarities influence the objectives of statistical data editing procedures, a short discussion about interactions between main objectives of E&I and peculiarities of administrative data follows.

Objective OB1

The identification of source of errors becomes in this context particularly important. In fact, one of the main problems is that the definition of collected variables is not designed for the survey purposes, and even after a process of harmonisation, some differences may still remain. The process of editing can help to reveal unexpected differences and to find whether there is a systematic nature of the error suggesting that the definitions are still not completely harmonised. Unfortunately, the improvement of the statistical process is limited by the fact that the process is not completely under the control of the NSI. Most of the times it is not easy or even impossible to return to the administrative entity collecting data and to make the agency change the definition of the variables, the data collection and so on.

Objective OB2

As for E&I of survey data, the data quality assessment in terms of input and output data is a key aspect also for statistics based on administrative data. The fact that two separate entities influence the data and the data production process, i.e., data holder and statistics provider (NSI), implies that two different points of view can be used for quality evaluation: a data perspective and a perspective oriented to the production of statistics. The first one is useful to provide information to the data holder to improve data quality for other data collection occasions, while the second one is important to measure the quality of the statistics provided inside and outside the NSI.

Objective OB3

The generally large dimension of databases has an impact on the detection and correction of influential data (which especially characterise quantitative variables), since for their treatment an expensive data editing procedure based mainly on re-contacting units is generally adopted. On the other hand, the use of multiple data sources may lead to have multiple observed values for a single observation, this information can be used to improve the selective editing procedure in terms of both identification of influential errors and value correction when an influential observation is selected. The same considerations hold when longitudinal information is available on units covered by administrative sources. These aspects will be later discussed in the subsection on editing methods.

Objective OB4

In case of integration of several data sources, the data consistency becomes an essential aspect, because the integration will increase the possible conflicts into the available information. However, as previously stated, the presence of multiple observations is an important aspect that can improve the E&I procedures, although at this time not many methods are developed to exploit as much as possible this richness of information. This issue will be discussed in the subsection on editing methods.

In the end, we can state that the general setting designed by the objectives of E&I of survey data remains still valid for administrative data. On the other hand, it is important to be aware of the impact of peculiarities of administrative data giving a different perspective to the objectives, those peculiarities will have an impact in the design and use of methods for E&I of administrative data

2.2 Types of errors in administrative data

As previously discussed, also in case of administrative data, one of the most important objectives of statistical data editing is to deal with errors, for this reason is important to discuss the characteristics of errors affecting administrative data. Before starting with the description of errors is useful to clarify a question: are administrative data affected by errors? It is difficult to imagine that data relating, for instance, to tax declaration can be affected by errors. It is nowadays accepted the idea that administrative data can be affected by errors (Groen, 2012), in fact also for this type of source errors may arise in many phases of the data production process, e.g., at the data transmission phase between data holder and NSI. Furthermore, there are also less controlled administrative data sources where the information is not so immediately sensible to make the data holder perform a check. A discussion about errors can be found later in this section.

Summarising, as well as survey data, administrative data are normally affected by different types of error: in the most recent literature, it is actually accepted that the non-sampling errors that normally emerge in surveys may also occur in registers (Bakker, 2011; Zhang, 2012). We start from the assumption that all the errors dealt with at the E&I phase in case of a single source survey are potentially present in a single administrative data source, hence the discussion is focused on the new additional aspects characterising errors in administrative data, with special attention to the case of statistics produced by integrating different data sources.

The E&I procedures are mainly designed to deal with measurement errors and missing values, the latter concerning usually item non-response. These sets of errors are analysed in the following.

Measurement errors are defined as differences between the recorded values of variables and the corresponding real values (*intended measure* of the variable). They mainly arise because of the fact that administrative sources are the result of processes which, being designed for purposes other than statistical, may use different concepts and/or definitions than those required for the specific statistical purposes. Important differences between the sources of measurement errors in survey data and in administrative data derive from the fact that the measurement process is very different in the two situations. In surveys using questionnaires, measurement errors derive from a cognitive process (comprehension of the question, retrieval of the information, judgment and estimation, reporting the answer) which also acts in case of administrative data but is not the most important one. A most important role in this case is played by administrative and legislation rules and accounting principles (Wallgren and Wallgren, 2007, p. 180). Typical measurement errors in administrative data are errors in accounting routines, or misunderstanding due to legally complicated questions, or errors deriving from the misspecification of rules used for deriving statistical variables from administrative variables. Furthermore, as some variables recorded for administrative purposes are more important than others, their accuracy is expected to be superior, as it can be assumed that enterprises answer to less important questions with lower precision. It is worth mentioning that the cognitive process also acts in case of administrative data: measurement errors may derive from the fact that respondents may provide different data to the different government agencies depending on their specific purpose, they may understand administrative concepts and definitions incorrectly (thus introducing errors by deviating from definitions, e.g., including wrong elements in the reported variables), or they can make unintentional errors in providing information.

Among measurement errors, also in case of administrative data variable values may contain ***systematic errors*** (cf. "Statistical Data Editing – Main Module"), which in this case can be due, for example, to a misinterpretation of record descriptions, originated by changes in the record descriptions and/or variable names in the administrative data bases.

An important source of errors for statistics based on multi source administrative data is the process of data integration itself. When the statistical population is created, objects are adjoined and linked, variables are imported from different sources and derived variables are created. The most relevant types of errors associated to the integration process are *coverage errors, identification errors, consistency errors, aggregation errors, missing values* (Zhang, 2012; Wallgren and Wallgren, 2007, p. 177). While coverage errors are not usually treated through E&I, the others are dealt with by or have an impact on the E&I process, for this reason they are described in the following.

Identification errors. They may be originated by errors in identifying variables used to match the different sources. As a consequence, identification errors may give rise to doublets, mismatches (e.g., false hits), item and total non-response, data inconsistencies (as variables may be referred to not properly matched objects). Identification errors may also generate outliers, and influential errors.

Consistency errors. They may also originate from the integration of variables from many sources. This type of error is especially increased when using multi source data, on the contrary with a single statistical survey, the use of a unique questionnaire ensures a better consistency in the data. Consistency errors can be caused by errors in units and errors in variables. They may also have a longitudinal origin, e.g., due to identifying variables either in error or changing over time for a same unit, splits/fusions of a unit over time.

Incoherent variable values giving rise to consistency errors in microdata may occur in the situation where the integrated administrative sources are overlapping regarding (a subset of) variables.

Inconsistencies with information from other sources and outliers can be originated from modifications of the variables' definitions adopted in a source (e.g., resulting from legislative changes), and from the fact that units may change their structural characteristics (e.g., fusions or splits). Outliers can also be determined by taxation measures that produce anomalous changes in variables values over time, and by integration errors (e.g., different units are linked in administrative sources). Outliers can either correspond or not to influential errors, depending on their impact on the target estimates.

Aggregation errors. They may occur when data from different administrative sources with different types of units are integrated in order to derive statistical variables (Wallgren and Wallgren, 2007), e.g., enterprise labour cost deriving from fiscal archives on enterprise employees. Aggregation errors may originate internal inconsistencies among variables referring to the same unit, outliers and longitudinal inconsistencies.

Missing values. As for statistical surveys, also in case of administrative data, missing values may correspond to two types of non-response : *unit non-response* (all the information for a statistical unit is unavailable) and *item non-response* (incompleteness of information, for some units, on topics which are of interest for statistical purposes). In case of administrative data, unit non-response corresponds to under-coverage, for example, when the integrated administrative sources relate to sub-populations which do not cover the overall target population. Item non-responses typically derive from the fact that the content of administrative sources is defined on the basis of administrative requirements, thus not all topics of interest may be covered by the administrative data. Possible sources of item non-response can arise for other different reasons: variable values can be missing for certain objects due to flaws of a source; mismatches at the integration phase due to missing objects in a source, giving rise to missing values for all the variables which are imported from that source; reported values which are "cancelled" as recognised invalid at the editing stage; values which fail to be reported, or are reported with a delay. Item non-response can also be associated to the fact that the content of a source is subject to modifications, resulting from legislative changes, like the drop-out of some information from the administrative forms; in a longitudinal perspective, non-responses can also appear as missing information on target variables for units considered over time: this can be due again to modifications of the units (fusions/splits, other structural changes) or to changes in legislation. Finally, as administrative sources may refer to either a point in time (i.e., they describe the units set at that point in time), or to a calendar year (in this case they contain all units that have existed at any point during the year), item non-responses may rise when sources with different time characteristics are integrated.

2.3 Data editing methods for administrative data

In this section we focus the attention on methods which can be used to detect and treat measurement errors and item non-response, that are in fact the errors dealt with by an E&I procedure..

Several classifications for the data editing techniques are available; we follow the one proposed in "Statistical Data Editing – Main Module". The techniques can be classified as:

1. Deductive editing.
2. Selective editing.
3. Automatic editing.
4. Interactive editing.
5. Macro-editing.

The order follows the strategy that is generally adopted in an E&I process for a statistical survey (cf. "Statistical Data Editing – Main Module").

In this section we discuss the impact of the peculiarities of administrative data on the features of each data editing technique.

Deductive editing is the phase where methods for detecting and treating errors with a structural cause that occurs frequently in responding units (systematic errors) are used (see "Statistical Data Editing – Deductive Editing"). In administrative data, especially when more sources are used, deductive editing has an important role in the production process. Variables collected in the administrative sources may have similar definitions but they may have structural gaps given to the convenience of declaring some information in an item rather than in another one, for instance, declaring something either in a cost or in an investment item. The first step in an E&I process should be to look for systematic errors in the observed values, also in the case the definition of variables is almost the same with respect to the corresponding statistical target variable. Hence, deductive editing is substantially the same as the one carried out in a classical data editing process, in fact the detection of systematic errors implies the involvement of subject matter experts, and the error treatment, that is usually completely automated, is not affected by the large dimension of administrative databases.

The aim of ***selective editing*** is indeed the optimisation of the process of selection of units to be deeply revised (in most cases, re-contacted) by restricting the editing only to those affected by an important error, and this naturally stresses the importance of selective editing in this context where data sets have usually a large dimension. On the other hand the use of selective editing is actually limited by resources' constraints because even a small percentage of units to be analysed may be too large in a large data set. A further constraint for selective editing on administrative data derives from the difficulty of re-contacting units for this kind of data. This limitation is alleviated when multi-source data are used, in this case the availability of different values for the same observation is an important aspect that can help the statistician in understanding where the error is located and to recover a likely value. The previous considerations mainly illustrate the problems in applying selective editing to administrative data. However, some further remarks concerning positive aspects of selective editing with administrative data are worthwhile to be mentioned. In selective editing, observations are prioritised according to a score function measuring the impact on the target estimates of the expected error in the unit. The error is frequently measured by comparing the observed value with a suitable prediction. In the context of administrative data, there is frequently the possibility of using longitudinal data, and this can improve the efficiency of selective editing as better predictions can be obtained. Finally, it is worthwhile to note another specific difference characterising the application of selective editing in administrative data with respect to the survey data. In a survey, the error is generally weighted with sampling weights. Since the prioritisation of an observation should be based on the impact of the error on the estimates, the final sampling weights should be taken into account in this process. In practice, this can be rarely performed, as final weights are generally computed once the editing step is completed, so an approximation is generally used by considering initial sampling weights. In the case of administrative data this problem is naturally overcome because sampling weights are not an issue for these kinds of data and a more precise estimation of the impact of errors on estimates can be obtained.

Automatic editing refers to all E&I procedures that detect and treat errors and missing values in a data file in a fully automated manner, i.e., without human intervention (see "Statistical Data Editing – Automatic Editing"). In the last years, most of the methods for automatic editing are based on the Fellegi-Holt paradigm, which means that the smallest number of fields should be changed to a unit to be imputed consistently. The algorithms are based on edits that represent rules/constraints characterising the relationships among variables.

In principle, if the focus is just on one data source, we are in the same situation as the one we would have in an E&I process of statistical survey data. However, as already remarked, most of the times different data sources are integrated, and in this case some additional problems may arise. A first issue to take into account is whether the data sources should be treated simultaneously as a unique data set after the integration process. This could be an interesting option, because the amount of information would increase, and an improvement in the E&I procedure is expected. In this case, edits simultaneously involving variables of the different data sources should be considered. A special but not infrequent case is when the same (at least in principle) variable is observed in the different data sources. For the sake of simplicity, let us suppose that there are only two data sets with the same variable. According to the Fellegi-Holt approach, we are assuming that with a high probability at least one of the two variables in turn is not affected by error. In the case that this assumption is not reliable, a different approach should be followed, for instance, a prediction conditionally on the observed values of the two variables can be obtained. Techniques developed to this aim are described in the module "Micro-Fusion – Reconciling Conflicting Microdata".

Concerning ***interactive editing*** for administrative data, the most relevant aspect is that, as already remarked, it is frequently not possible to re-contact the observed units, so one of the main advantages motivating interactive editing declines. However, interactive editing can be considered effective in order to understand error sources and possibly resolve errors in the short term, while in the long term it can contribute to the increase of the subject-matter expertise for the staff working on administrative data, increasing their knowledge of the characteristics and the contents of administrative data and gaining understanding of how the data can be used in a more suitable way (Wallgren and Wallgren, 2007).

Macro-editing aims at looking for anomalous aggregates. The anomalies are identified based on the comparison of aggregates with some reference values that, for instance, may be obtained by previous published figures. Once anomalous aggregates are selected, a drill-down procedure is applied in order to find the units that mostly contribute to this behaviour (see "Statistical Data Editing – Macro-Editing"). This editing approach requires the computation of the final aggregates (e.g., domain estimates), and for this reason, in the usual E&I procedure it is generally performed at the end of the E&I process. In this context, one generally works on complete data sets, in fact administrative data are gathered for other purposes and they are usually provided to the NSIs at the end of their collection. This implies that in this context macro-editing methods can be used at the beginning of an E&I procedure in order to look for important errors.

Macro-editing can be a useful tool to reveal whether some important errors due to an incomparability of the sources in some estimation domain are still present in data. For instance, it can happen that the definition of a variable is the same in two data sources. Nevertheless, for a specific economic sector some particular businesses could not provide the complete amount of the value in one source because of fiscal benefits typically allowed only for that segment of units. Macro-editing can be useful to isolate those critical situations that the subject matter expert may study and interpret in order to fix the problem wherever it is possible. Macro-editing can also reveal errors due to data linking or to the incomplete delivery of some sources, as anomalous aggregates may result from not enough covered domains from one time period to the subsequent one.

As already mentioned, administrative data are subject to partial non-response as well. ***Imputation*** (see the topic "Imputation") can be used to manage missing values in order to obtain a completed data set on which the usual statistical analysis can be applied. The methods usually adopted are based on the missing at random (MAR) assumption that is, roughly speaking, the probability of non-response on a given variable depends on the observed values and not on the unobserved ones of the variable itself. For instance, missing values in administrative data can be due to lack of timeliness, and it is generally supposed that businesses answering in due time have the same behaviour as the not observed ones. Actually this situation could hide the presence of a problem in the business, and in this case the estimates could be biased because the observed and non-observed populations are actually different. A similar concept applies in the case of an integrated use of administrative data. It can happen that each administrative source covers only some specific part of the target population. Imputation can be used to complete the missing values, again under the assumption that the population not covered has the same behaviour of the observed one.

Finally, since the production process of administrative data is generally beyond the control of NSIs, a continuous assessment of the data quality should be planned. Edit rules and macro-editing based approaches could be used to this aim. An anomalous rate of edit failure and/or anomalous variation of statistical aggregates in two consecutive times could alert data producer that some important changes could have been introduced in the administrative data production process, which could be related to a change in the data collection, to a change in the legislation that impacts on the definition of measured variables, consequences of a different fiscal policy, and so on.

2.4 Information about data quality

One of the main goal of E&I is to provide information about the quality of the data collected and published.

Quality of statistical output has several dimensions, they are thoroughly discussed in Eurostat (2011) for the European Statistics Code of practice, Eurostat (2009) for a handbook (soon to be revised) on reporting quality of statistical data according to the European output quality components, and the handbook module “Quality Aspects – Quality of Statistics”.

In this section it is important to refer to the quality dimensions in the context of administrative data in order to describe on which of them the E&I is a useful tool for providing information. In the BLUE-ETS (2011) document, the quality dimensions of administrative sources and the related indicators are discussed. In that document the focus is on the quality dimensions of the administrative data sources in the input phase of a statistical production process, this point of view is adopted in this paper as well. As far as the quality dimension of the statistical output based on administrative data is concerned, we assume that at the end of the E&I process data are statistically transformed, and hence the general considerations made for statistical output based on survey data are still valid. This is a simplistic position, that is also motivated by the fact that at this time this issue is still under discussion, and further studies are needed in this context. For the use of E&I procedures as a useful tool for providing information on quality of statistical data, the reader may refer to EDIMBUS (2007).

A first interesting remark relates to the point of view chosen to look at the quality aspects. It reflects the peculiarity of statistics based on administrative data where generally two different main actors are involved: the data holder and the statistics provider (NSI). Two main points of view are introduced: a data archive perspective and a perspective oriented to the production of statistics. In the first one, the quality is independent of the specific statistical use of the administrative data that is supposed to be done, while in the second one the quality is related to the statistical use of the data planned at the NSI. Both these aspects are important for E&I, in fact the first one has to be assessed in order to foster data holder to improve the quality of the data, while the second one is related to the quality of published data.

In the BLUE-ETS document, the following quality dimensions are defined:

1. *Technical checks*, that is the technical usability of the file and data in the file.
2. *Accuracy*, that is the extent to which data are correct, reliable, and certified.
3. *Completeness*, that is the degree to which a data source includes data describing the corresponding set of real-world objects and variables.
4. *Time-related dimension*, in which timeliness, punctuality, and overall time lag applied to the delivery of the input data are taken into account.
5. *Integrability*, that is the extent to which the data source is capable of undergoing integration or of being integrated.

The *technical check* dimension is mainly related to IT aspects, e.g., data accessibility, correct conversion of the data, data complies with the metadata-definition. These aspects are not related to an E&I procedure as it is defined in “Statistical Data Editing – Main Module”.

E&I has certainly impact on *accuracy*, and it naturally provides information about some dimension indicators described in BLUE-ETS (2011) related to this aspect. Some of the dimension indicators for accuracy proposed in BLUE-ETS are supposed to measure:

- *Measurement error*: deviation of actual data value from ideal error-free measurement;
- *Inconsistent values*: extent of inconsistent combinations of variable values;
- *Dubious values*: presence of (or combinations of) implausible values for variables.

Those elements are treated and analysed during an E&I procedure, and indicators measuring them are developed and generally automatically provided by the usual procedures (see EDIMBUS, 2007).

E&I may be useful to gather information also for other quality dimensions, that apparently are less naturally related.

Completeness is a concept referred to units and variables, and for the latter the quality dimension indicators proposed in BLUE-ETS (2010) are: the amount of missing values and the amount of imputed values. As previously stated, the treatment of missing data (imputation) is one of the main activities carried out in an E&I process; hence, indicators on those aspects are easily obtained in this context.

As far as the *time related dimension* is concerned, a proposed indicator focuses on the stability of variables. To this aim, the comparison in different times of indicators generally provided by E&I may be useful: for instance, an anomalous variation of the failure rates of some edits may hide some changes in the administrative data production process or in the source contents, or in the use of a different definition for a variable, or in a different data collection mode. Also the comparison of the amount of imputed values and missing data can reveal some changes in the data source which have to be taken into account in order to avoid biasing effects on statistical results.

A summary of the editing undertaken and the results of the checks should be sent to the database owner to make him aware of the problems possibly existing in the data set, in order to reduce them as much as possible in the future and improve the overall quality of the data. As a consequence, managing and improving co-operation with administrative bodies plays a central role in this context:

NSIs need to increase co-operation and to determine appropriate incentives in order to improve the overall communication and interaction with data owners, to get them to set up better editing practices and conform to statistical classifications and definitions, and to provide feedback to the NSI in the data verification process (Shlomo and Luzi, 2004).

3. Design issues

In the design of an E&I process for administrative data the first important issue to take into account is whether the target statistics are based only on a single administrative source or on the use of multiple integrated administrative sources. Moreover, editing strategies must take into account the trade-off between the potential gain in accuracy deriving from the availability of detailed and extensive information, and the additional costs needed for validating it.

When only one source is used, as discussed in the previous sections, we are in a similar situation to that of E&I of a single survey, even if we remind that peculiarities of administrative data should be taken into account because of their impact in the E&I methods. The reference flow-chart introduced in the module “Statistical Data Editing – Main Module” can be applied to this case.

When more sources are integrated, different scenarios can be depicted.

A first scenario may consist of the following macro-phases:

1. check separately each single administrative source;
2. integrate the edited data sources;
3. edit the integrated sources in order to assess the consistency among variables’ values obtained from the different sources.

This is actually the flow-chart reported in Wallgren and Wallgren (2007, p. 101).

The drawback of this way of proceeding is that it is resource demanding since many different E&I procedures must be set and applied, and it is well known that the E&I is one of the most expensive parts of the statistical production process. Moreover, not all the amount of information is used at the same time, for instance, for the imputation of a variable in a data source it could be useful to exploit variables observed in the other data sources. Let us imagine the case when two data sources are integrated and in one source the income is observed, while in the other one information on consumption is gathered. The imputation of the two variables separately would disregard the strong relationship existing between them. An advantage of this way of proceeding is that certain typologies of errors (e.g., systematic errors like unity measure errors, balance errors, errors due to incomplete delivery of data for some administrative objects) can be removed from each single source before the integration phase, thus reducing the amount of consistency errors on the linked data deriving from these situations; longitudinal information could be used at this stage.

An alternative scenario corresponding to the opposite solution is:

1. integrate the sources;
2. apply an E&I procedure to the integrated data set.

In this case, less resources would be demanded since only one data verification process is required, but the complexity of such a process would increase. Furthermore, as the integrated data set is not generally composed of all the variables observed in the different administrative sources, in this case some relations linking variables in each data source could be disregarded.

A third scenario is a compromise of the previous ones:

1. apply a 'light' E&I procedure to each single administrative source;
2. integrate the edited data sources;
3. edit the integrated data sources.

The question is when an E&I procedure can be defined as light. The idea is that the time and effort spent in editing sources should be minimised while maintaining an acceptable level of quality of the data sources. This general idea resembles what is done in selective editing, where the effort is focused on the most important errors having a high impact on the target aggregates. This situation is slightly different because there is no requirement on a sufficient level of quality of aggregates for each single data source, but the level of quality is required at micro level: in effect, the use of each single source will be in a micro perspective given that the integration process is generally performed at this level. A proposal could be that of applying only corrections of systematic errors in the first editing step.

It is clear that a general flow-chart is not available; however, at least three scenarios have been designed. The main point is to see the E&I process as a unique process possibly composed of two steps. The choice of the most appropriate strategy should be based on the trade-off between the expected quality of the final aggregates and the resources which are actually available to obtain the required level of quality. Concerning the latter, an element which can be considered as relevant to increase the effectiveness of editing and correction activities is the availability of subject-matter experts, who are familiar with the administrative systems that have generated the data and their specific contents, and who are in good relations with the data providers.

Finally, independently on the chosen scenario, indicators providing information about input and output data quality should be part of the E&I process. Moreover, since the process of gathering information is out of control of NSIs, it is important to establish a system of indicators alerting about some possible changes in the data production process of the data holder, in order to avoid important and non-measurable errors in the published statistics.

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Bakker, B. F. M. (2011), Micro-Integration: State of the art. In: *Report WP1: State-of-the-art on Statistical Methodologies for Data Integration*, ESSNET on Data Integration, available at http://www.cros-portal.eu/content/wp1-state-art.

BLUE-ETS Project (2011), *Deliverable 4.2: Report on methods preferred for the quality indicators of administrative data sources*.
Available at http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf.

EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

Eurostat (2009), *ESS Handbook for Quality Reports*. Eurostat Methodologies and Working papers.

Eurostat (2011), *European Statistics Code of Practice*. For the national and community statistical authorities. Adopted by the European Statistical System Committee 28th September 2011 (revised version).

Groen, J. A. (2012), Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *Journal of Official Statistics* **28**, 173–198.

Nordbotten, S. (2010), The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries. In: Carlson, Nyquist, and Villani (eds.), *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn*, 205–225. Available at officialstatistics.wordpress.com.

Shlomo, N. and Luzi, O. (2004), Editing by Respondents and Data Suppliers. In: *Federal Committee on Statistical Methodology, Statistical Policy Working Paper 38: Summary Report on the FCSM-GSS Workshop on Web-based Data Collection, April 2004*, 75–90.

Statistics Canada (2010), *Survey Methods and Practices*. Catalogue no. 12-587-X.
http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.pdf.

Wallgren, A. and Wallgren, B. (2007), *Register-based statistics – Administrative data for statistical purposes*. John Wiley and Sons, Chichester.

Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica* **66**, 41–63.

Interconnections with other modules

8. Related themes described in other modules

1. Micro-Fusion – Main Module
2. Statistical Data Editing – Main Module
3. Statistical Data Editing – Selective Editing
4. Statistical Data Editing – Macro-Editing
5. Imputation – Main Module
6. Weighting and Estimation – Estimation with Administrative Data
7. Quality Aspects – Quality of Statistics

9. Methods explicitly referred to in this module

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Statistical Data Editing – Deductive Editing
3. Statistical Data Editing – Automatic Editing
4. Statistical Data Editing – Manual Editing

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

14. Module code

Statistical Data Editing-T-Administrative Data

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	13-03-2013	first version	M. Di Zio, O. Luzi	Istat
0.2	17-06-2013	introduction of a new section concerning quality indicators	M. Di Zio, O. Luzi	Istat
0.3	07-08-2013	minor revisions	M. Di Zio, O. Luzi	Istat
0.3.1	04-10-2013	preliminary release		
0.4	20-12-2013	revision based on EB comments	M. Di Zio, O. Luzi	Istat
0.4.1	09-01-2014	revision based on EB comments	M. Di Zio, O. Luzi	Istat
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:13