Estimation methods for the integration of administrative sources

Task 2: Identification and description of all the possible statistical tasks where using estimation methods can be envisaged in order to integrate administrative sources for a given use

Contract number:	Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252
Responsible person at Commission:	Fabrice Gras Eurostat – Unit B1
Subject:	**Deliverable D2**
Date of first version:	28.09.2016
Version:	V3
Date of updated version:	09.01.2017.
Written by :	Marco Di Zio, Ton de Waal, Sander Scholtus, Arnout van Delden, Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto, Li-Chun Zhang
Sogeti Luxembourg S.A.	Laurent Jacquet (project manager)
	Sanja Vujackov

Table of contents:

1 Introduction

The ultimate aim of this work is to discuss the possible tasks to be performed in order to use integrated administrative data and to relate them to the usages and the related estimation methods. To this aim, it is useful to identify and describe the possible statistical tasks where statistical estimation methods can be envisaged in order to integrate administrative sources for a given usage.

The mapping between usages, statistical tasks and statistical integration methods need to resort to some classification.

In order to describe the statistical tasks, we may refer to the classification introduced in the Generic Statistical Business Process Model (GSBPM v5.0) that aims at describing the steps for the production of statistics; see UN/ECE (2013).

This model allows to represent the statistical production process by means of building blocks representing the business processes to go through in a statistical production process. The peculiarities of each task with respect to a statistical production process based on integrated administrative data will be described.

2 Applying the GSBPM to the usages characterised by integrated administrative data

The GSBPM was developed by the UN/ECE and the Conference of European Statisticians Steering Group on Statistical Metadata. The GSBPM describes and defines the set of business processes needed to produce (official) statistics. It was introduced with the idea of providing a basis for statistical organisations to agree on standard terminology to aid their discussions on developing statistical metadata systems.

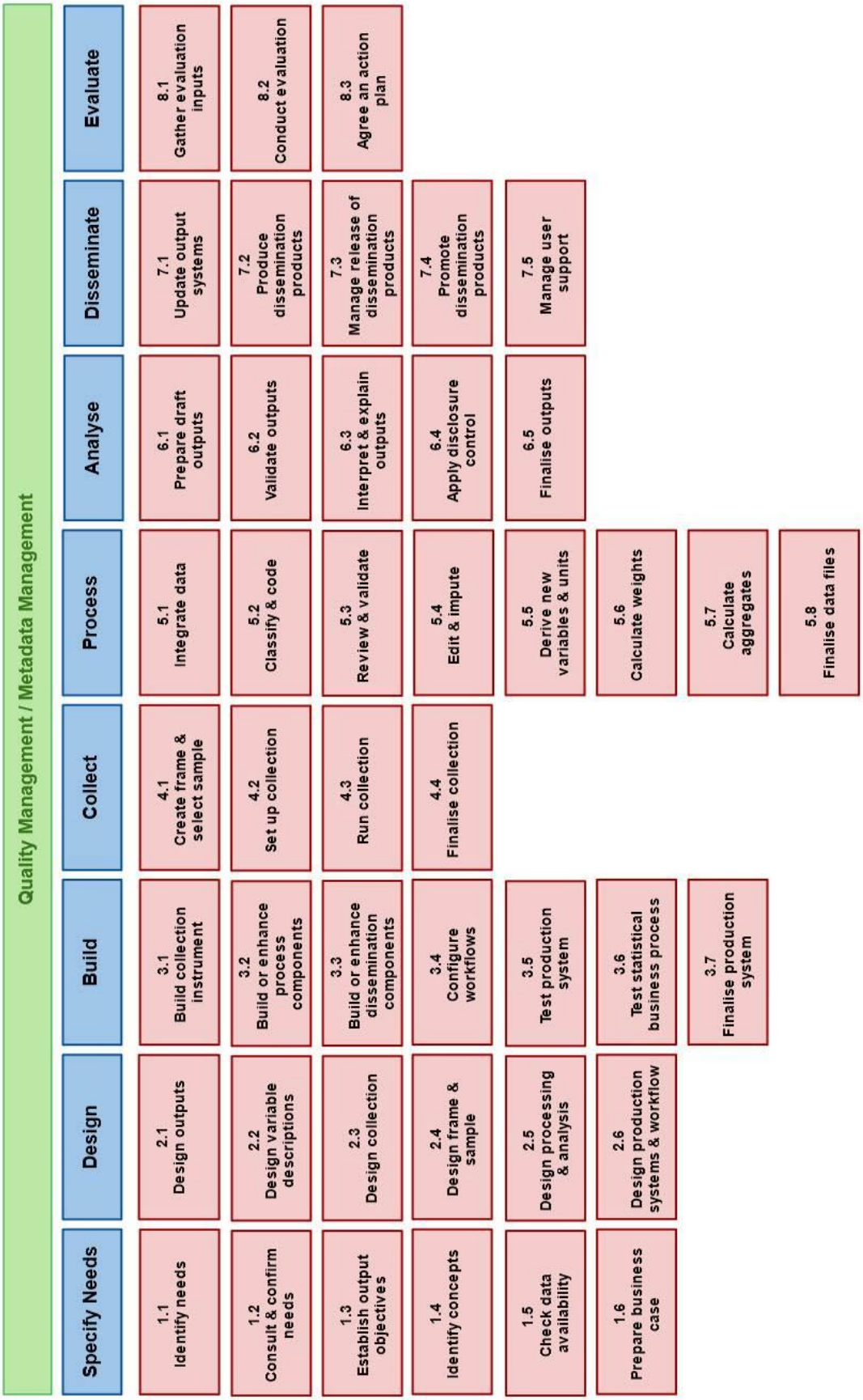
The GSBPM is intended to apply to all activities undertaken by producers of official statistics, at both the national and international levels, which result in data outputs. It is designed to be independent of the data source, so it can be used for the description and quality assessment of processes based on surveys, censuses, administrative records, and other non-statistical or mixed sources. The GSBPM comprises three levels:

- Level 0, the statistical business process;
- Level 1, the eight phases of the statistical business process;
- Level 2, the sub-processes within each phase.

A description of the sub-processes is given in UN/ECE (2013).

Levels 1 and 2 are represented in Figure 1. Although all the main phases to go through in a statistical production process are represented, GSBPM is not a rigid framework in which all steps must be followed in a strict order, instead it identifies the possible steps in the statistical business process, and the inter-dependencies between them. Hence, not all the tasks are necessarily covered in a production process, and loops between tasks may be used to represent the process.

Figure 1. Levels 1 and 2 of the Generic Statistical Business Process Model



In Task 1, the main usages of integrated administrative data are identified. In this section, a mapping between GSBPM and the main usages requiring integrated administrative data is provided. A reference paper studying the application of GSBPM to Business Register is UN/ECE (2011).

2.1 Direct Tabulation

This is the category where administrative data are used to produce statistics without resorting to any statistical data. Important cases are the register-based census-like statistics in a number of European countries, UN/ECE (2014).

Once we recall that the 'collect' level concerns all kinds of data, including administrative data, it is manifest that all the process steps, except the ones clearly referring to sample surveys, are relevant. In fact, administrative data enter the process in the 'collect' phase, and they go through all the possible steps. A particularly important sub-level for these data is 'Integrate data' (sub-process 5.1). In this sub-process, all the integration tasks needed for 'integrating' data are included, as GSBPM states:

"*This sub-process integrates data from one or more sources. It is where the results of sub-processes in the "Collect" phase are combined. The input data can be from a mixture of external or internal data sources, and a variety of collection modes, including extracts of administrative data. The result is a set of linked data. Data integration can include:*

- *combining data from multiple sources, as part of the creation of integrated statistics such as national accounts;*
- *matching / record linkage routines, with the aim of linking micro or macro data from different sources;*
- *prioritising, when two or more sources contain data for the same variable, with potentially different values*".

2.2 Substitution and Supplementation for Direct Collection *(Replacement for data collection)*

This is the category where administrative data are directly used as input observations for the production of statistics but are not sufficient for achieving all the objectives of the statistical program.

As in the previous case, once data enter the phase through the 'COLLECT' sub-process, all the other sub-processes are relevant. The difference with respect to the previous usage is that, since a sample survey is used, also the sub-processes concerning sampling are relevant, e.g. sub process 5.6 'calculate weights'. Similarly to the previous case, the '5.1 Integrate data' sub-process is still the central sub-process of the production of statistics under this setting.

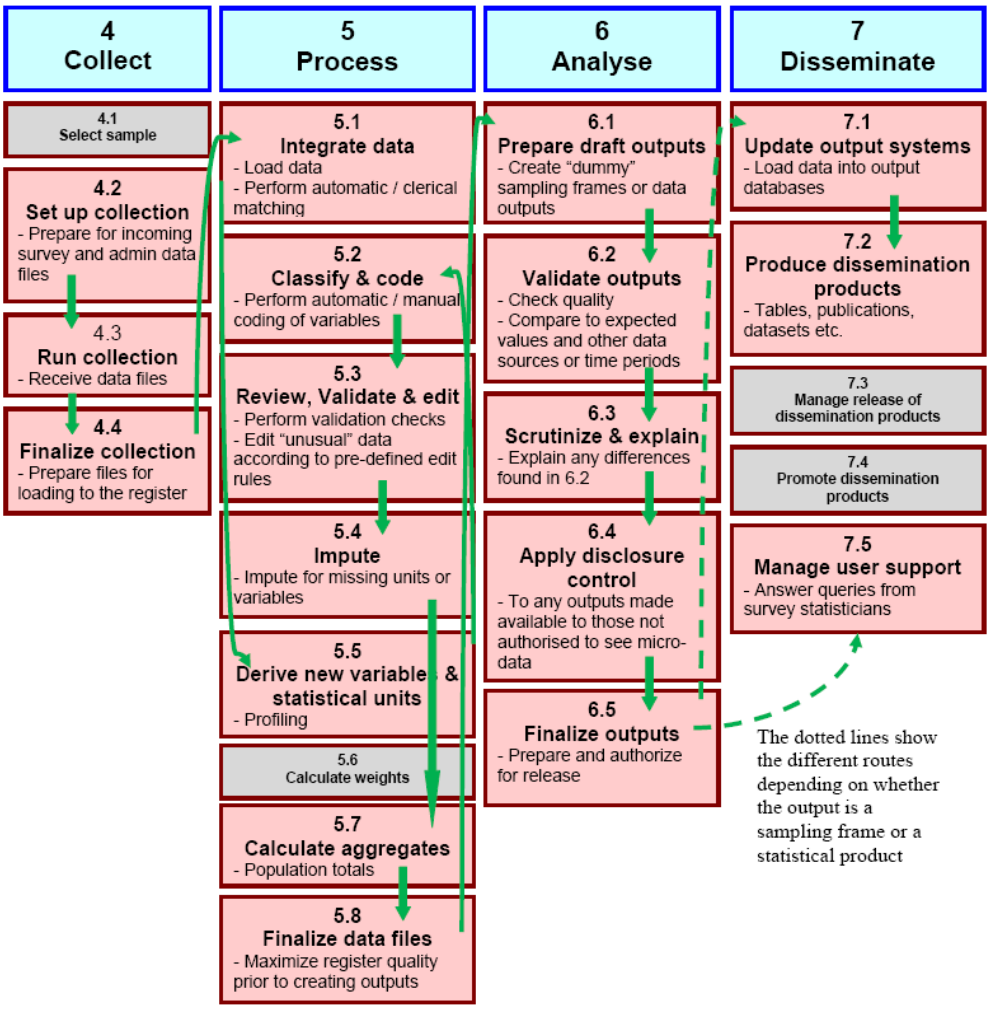
2.3 Creation and maintenance of registers and survey frames

The application of GSBPM to the creation and maintenance of a register is well discussed in UN/ECE (2011). As stated in the document describing GSBPM, the creation and maintenance of a register can be modelled by means of GSBPM, that is "*… the inputs are similar to those for statistical production (though typically with a greater focus on administrative data), and the outputs are typically frames or other data extractions, which are then used as inputs to other processes*".

The main characteristic of such a product is the continuous work that is necessary for maintaining; this can be different from some regular statistical outputs where the periodicity is not frequent. However, it does not affect the sub-process elements described in GSBPM. The sub-processes that are not (or less) relevant are those concerning sample surveys.

Figure 2 reports the illustration given in UN/ECE (2011) about the application of the GSBPM model to the Business Register, referring in particular to the phase 'Process'. In grey are the less relevant sub-processes. The arrows represent a possible path to go through the phases. It is, however, worthwhile to remark that given the nature of GSBPM, other paths can be imagined.

Figure 2. The application of the GSBPM model to the Business Register



2.4 Editing and imputation

Administrative data can be used to check and impute survey data. The specific usages of editing and imputation are quite broad, in fact administrative data can be used to build edits (rules for checking consistency), they can be used for a comparison with actual data, and so on. Despite the variety of specific applications, this usage refers to two specific sub-processes of GSBPM that are "review and validate" (sub-process 5.3) and "edit and impute" (sub-process 5.4).

Other phases of the GSBPM can be indirectly involved, depending on the specific usages. For instance, if a check is performed at micro-level, this probably requires that data are linked, and the sub-process 5.1 'Integrate' is needed. On the other hand, if administrative data are used to impute missing data, the sub-process 5.5 'Derive new variable' may be needed.

2.5 Indirect estimation

In this category, administrative data are one of the inputs of the estimation process. It means that they are either useful or necessary for producing the estimates, but they do not constitute all the input data.

The most relevant sub-process is 'calculate aggregates' (sub-process 5.7), that is the part of GSBPM where the estimation is performed. Of course, also in this case, other processes may be indirectly involved. They are concerned with all the tasks that make the administrative data usable for the data at hand, so for instance the sub-process 'integrate' may be needed.

2.6 Data validation/confrontation

For this usage, administrative data are exploited to validate data both in terms of macro and micro levels. The relevant phases of GSBPM are the sub-processes 5.3 ('Review and validate') and 6.2 ('validate outputs'). Similarly to the previous descriptions, a task of integration may be needed.

2.7 Conclusions on applying the GSBPM to relate usages, statistical tasks and statistical integration methods

We notice that if administrative data are used as input data, mostly all the GSBPM sub-processes starting from the 'COLLECT' level are relevant, otherwise administrative data mainly refer to specific sub-processes; in all cases, an overarching sub-process as 'integrate' is generally needed.

GSBPM is necessarily general to describe all the statistical production process. It is useful to specialize the GSBPM with respect to the problem of using integrated administrative data, that is to indicate the statistical tasks for the statistical production with integrated administrative data and to classify them with respect to GSBPM.

3 Statistical tasks for using integrated administrative data

The statistical tasks we classify below can be considered as 'building blocks' of a process designed to enable and ease the usage of administrative data into a statistical production process. We notice that this definition includes also a statistical production process based only on a single administrative data source.

In general the statistical tasks refer to:

1. Transformation of administrative data in order to be used for our scopes, i.e. transforming objects and attribute in statistical units and measurements needed for the target statistics
2. Integration of the different data sources to join initially separate statistical information.

The statistical tasks are given in the following list:

I. Data editing and imputation (GSBPM 5.3, 5.4)

II. Creation of joint statistical micro data (GSBPM 5.1)
   a) Data linkage: Identification of the set of unique units residing in multiple datasets
   b) Statistical matching: Inference of joint distribution based on marginal observations

III. Alignment of statistical data (GSBPM 5.1, 5.2, 5.5)
   a) Alignment of units: Harmonisation of relevant units, creation of target statistical units
   b) Alignment of measurements: Harmonisation of relevant variables, derivation of target statistical variable

IV. Multisource estimation at aggregated level (GSPBM 5.1, 5.6, 5.7)
   a) Population size estimation: multiple lists with imperfect coverage of target population
   b) Univalent estimation: numerical and statistical consistent estimation of common variables
   c) Coherent estimation: aggregates that relate to each other in terms of accounting equations

Tasks II-IV are particular of a statistical production process based on integrated data. They refer to the integration problems mentioned in deliverable 1 named as 'problem of univalency', 'harmonisation of different data sources', 'linking different data sources' and the estimation based on integrated data. These arise chiefly for two reasons: reuse of administrative data that are not generated for statistical purposes, combination of multiple datasets that potentially overlap in units and variables.

Task I is not peculiar but nevertheless necessary of a statistical production process based on integrated data. It can take on special characteristics in the context of integrated administrative data.

The ordering of the list does not imply the sequence of the tasks in a statistical production process. In fact it is possible to have different combinations. For instance, some typical sequences are:

1. (I) → (II.a) → (III) → (IV)

   In this sequence, firstly data cleaning is carried out on each data source, then they are combined by means of record linkage, afterwards the integrated data are processed through micro-integration and aggregated estimation.

2. (I) → (III) → (II.a) → (I) → (III) → (IV)

   In this sequence, firstly each data source is cleaned and harmonised first, e.g. because each may have its own multiple input sources, then data sources are combined by record linkage, and finally data cleaning, micro-integration and estimation are performed on the integrated data.

3. (II.a) → (III) → (I) → (IV.a) → (IV.b)

   In this sequence, firstly all the datasets are brought together by record linkage, then unit and variable harmonization are carried out before data cleaning, and finally the target population size is estimated before obtaining various estimates that are statistically and numerically consistent.

We notice that tasks (I) and (III) can be performed on each input dataset before task (II), but also afterwards and based on the combined data (see Memobust, 2014a). Notice also that task IIIa can also be part of the construction of a base register (population register, business register or immobilisation register) that contains relations between administrative and statistical unit types.

3.1 Statistical tasks description

I. Data editing and imputation

Errors are virtually always present in the data files used by producers of statistics, also when data originate from external data sources (Groen, 2012). In order to produce statistical output of sufficient quality, it is important to detect and treat these errors.

Data editing procedures are designed to deal with errors and missing values. Errors are defined as differences between the recorded values of variables and the corresponding real values (intended measure of the variable) (Memobust, 2014). Missing values typically derive from the fact that the content of administrative sources is defined on the basis of administrative routines, thus not all population elements may be covered by the administrative data. For example, the variable highest education level may be missing for immigrants, although it is available for all the natives. Missing values can arise for other different reasons, among them: mismatches at the integration phase due to missing objects in a source, such that all the variables which could have been imported from that source are missing; reported values which are "cancelled" as invalid during editing;

values which fail to be reported, or are reported with a delay. The peculiarities of errors and data editing techniques for cleaning administrative data are discussed in Memobust (2014a).

When using multisource data, inconsistencies at micro level are frequently encountered. Integration of data sources at micro-level may give rise to *composite records* that consist of a combination of values obtained from different sources: for instance a register combined with values obtained from a survey for the same units (obtained by record linkage), an integration of several surveys with non-overlapping units, in which case a unit from one source is matched with a similar (but not identical) unit from another source. In addition, records with values obtained from different sources can also arise as a consequence of item non-response and subsequent imputation in which case the two sources are the directly observed values versus the values generated by the imputation method. In all these cases the composition of a record by combining information obtained from different sources may lead to consistency problems because the information is conflicting in the sense that edit rules that involve variables obtained from the different sources are violated. The purpose of editing conflicting micro-data is to achieve numerical consistency in the first instance and, ultimately, statistical consistency of the resulting aggregates.

It is worth noting that editing conflicting micro-data is used to resolve apparent numerical inconsistencies between *values* of the same variable from different data sources (composite records). While variable harmonisation (later on described) is applied to situations where there exist *similar versions* of the *same target* variable, for instance due to different definitions across the sources.

II Creation of joint statistical micro data

II.a Data linkage: combining data belonging to the same units

When administrative data are used in a multisource context, it may be required to link the units of the different data sources. The integration of data at micro level with the main purpose of accurately recognize the same real-world entity at individual micro level, even when differently stored in sources of various types, is known as record linkage. At the end, the record linkage process creates a micro dataset where, for the units that are identified to reside in multiple sources, the corresponding separate observations are combined into joint statistical data. Record linkage is also referred to as object identification, record matching, entity matching, entity resolution, reference reconciliation. In case of the National Statistical Institutes (NSIs) the joint use of statistical and administrative sources is a product of a rationalization of all the available sources to reduce costs, response burden and, most of all, to enrich the information collected in order to produce high quality statistics. Depending on the confidentiality constraints and the context, linkage activity could include activity related to the building of synthetic and anonymised identifier through the use of hashing techniques.

II.b Statistical matching: inference of joint distribution from marginal observations

In statistical matching (SM), the ultimate aim is to provide estimates on variables observed in distinct data sources (characterized by not having common units) by exploiting the available common information (variables observed in both data sources). The objective of SM can be achieved through either a *micro* or a *macro* approach. In the micro approach SM aims at creating

a synthetic data source in which all the variables are available. In the macro approach the data sources are used to derive an estimate of the parameter of interest, e.g., contingency table. It is worthwhile noting that the macro approach to statistical matching may be also classified in the statistical task IV 'Multisource estimation at aggregated level' described later on this document.

III. Alignment of populations and measurements

III.a Alignment of populations: Unit harmonisation

When administrative data are used, we need to transform objects into statistical units referring to the target population. This is a mandatory requirement valid both for usages based on a single data source, and for cases based on multisource data. After a preliminary analysis aimed at verifying such a consistency, it is sometimes necessary to start a process of harmonisation or alignment of units. This means that the objects observed in the administrative data sources must be transformed into relevant possibly several types of units, and the target statistical unit may then need to be created on the basis of them. For example, to create *living household* as the ideal unit for household income statistics, one needs to make use of units such as person, family, residence address, study or workplace address, etc. The different units need first to be aligned with each other, in order to improve the accuracy of the living household created based on them. For another example, to create *local kind-of-activity unit* as the ideal statistical unit for National Account, one needs to make use of legal units, tax units, enterprise, etc. The identification of a local unit and assignment of NACE code is a process of creation in truth.

III.b Alignment of measurements: Variable harmonisation

When administrative data are used, we need to transform attributes of the objects into measurements referring to the concepts we would like to measure, that is the target variables. This is a mandatory requirement valid both for usages based on a single data source, and on multisource data. After a preliminary analysis aimed at verifying such a consistency, it is sometimes necessary to start a process of harmonisation or alignment of measurements.

When data from multiple sources are combined, differences in definition can occur between variables in different sources. In particular, variables in an administrative data source are defined according to the administrative purposes of the register owner. These definitions may differ from those of the target variables for statistical purposes. For example, a tax authority collects data of value-added tax (VAT) declarations from businesses which contain turnover values. Since the administrative purpose of these data is to levy taxes on turnover, the tax authorities will be interested only in the amount of turnover of each business that is derived from taxable economic activities. Depending on the specific tax regulations that apply, for some sectors these administrative turnover values will differ from the turnover values that a statistical institute needs: some economic activities that are relevant for economic statistics may be exempt from taxes, and vice versa.

Another example is when the variable 'age-group' is observed in two data sources with different groups. A task for obtaining a unique variable 'age' is needed. This apparently deterministic task can involve statistical estimation method (e.g., classification techniques) whenever a well-defined mapping is not known.

In case differences in variable definitions occur between data sources, these variables need to be harmonised during data integration. That is to say, for each unit in the integrated data set, the values of the target variable according to the desired definition need to be estimated from the observed values that are available.

IV. Multisource estimation at aggregated level

This task refers to the phase of production of estimates by using integrated multisource administrative data.

Estimation methods in this case must deal with the problem that data are not obtained according to a random sampling design, and that observations may refer to specific subsets of units of the population (under-coverage), or out-of-scope units (over-coverage). In addition, when integrated multisource data are used, specific problems may arise. They are concerned with the problem of consistency and coherence of estimates.

IV.a Population size estimation: multiple lists with imperfect coverage of target population

When each one of the multiple sources has imperfect coverage of the target population, including both under- and over-coverage, statistical methods for population size estimation are needed. The most common approaches relate to capture-recapture (CRC) methods. In estimating a population size based on several administrative sources, the misalignment between the scope of the administrative data and that of the statistician poses several methodological challenges and sets us apart from a classical capture-recapture setting. For instance it is often useful to develop methods taking into account the dependence among data sources, the fact that some data sources refer to a specific subpopulations, and that data may contain units out-of-the target population that are not deterministically identifiable.

IV.b Univalent estimation

When using a mix of administrative data sources and surveys to base estimates upon, obtaining one estimate for the same phenomenon may become problematic as for different (combinations of) variables data on different units, e.g. different persons, may be available.

This means that different estimates concerning the same variable may yield different results, if one does not take special precautions. For instance, if one uses a standard weighting approach to produce estimates, where one multiplies observed counts or values with surveys weights, one may get different estimates based on two samples, because different units and hence different survey weights are used in the two samples.

In principle, these differences are merely caused by "noise" in the data, such as sampling errors. So, in a strictly statistical sense, different estimates concerning the same variables are to be expected and are not a problem. However, different estimates would violate the one-figure (univalency) policy and form a problem for the users.

A similar problem affects estimates when a time component characterises data sources. A common problem of macro-integration arises when times series data of the same target variable exist with

different frequencies in different sources, e.g. quarterly (or yearly) administrative data vs. monthly survey data. The data of the lower frequency may be based on a larger or more reliable set of data, in which case one may fix these and adjust the frequency time series accordingly, where one would like to keep the adjustments (to de defined later) as small as possible.

IV.c Coherent estimation

The estimates involved in a system of accounting equations, such as the supply-and-use table for National Account, are often derived based on different sources, and initially do not satisfy the accounting equations directly. This is a problem of coherent estimation, whereby the initial estimates need to be adjusted in order to satisfy the accounting equations. The resulting final accounts are affected by both the initial estimators and the method of adjustment. Appropriate summary of the "accounting uncertainty" can thus point to the most effective improvements needed of the initial estimators, as well as help to choose the efficient method of adjustment.

4 Relationships between statistical tasks, GSBPM sub-processes, and usages

The previously described statistical tasks can be connected to the GSBPM sub-processes.

The first task of **data editing and imputation** is connected to the GSBPM sub-processes 5.3 "Review and validate", and 5.4 "Edit and impute". These sub-processes include all the elaborations needed for data cleaning.

Unit harmonisation is related to the sub-process 5.5 "Derive new variables and units". In this sub-process data for variables and units that are not explicitly provided in the collection, but are needed to deliver the required outputs, are derived. New units may be derived by aggregating or splitting data for collection units, or by various other estimation methods. Examples include deriving households where the collection of observed units are persons, or enterprises where the collection of observed units are legal units.

Variable harmonisation is related to the GSBPM sub-processes 5.1 "Integrate data", 5.2 "Classify and Code", 5.5 "Derive new variables and units".

The sub-process "Integrate data" is referred to the integration of data from one or more sources. In UN/ECE (2013) it is explicitly mentioned one of the possible approaches used to obtain one value for a variable in the case the variables is observed in many sources: *prioritising, when two or more sources contain data for the same variable, with potentially different values.*

"Classify and Code" and "Derive new variables and units" are mainly related to the transformation of attributes related to administrative objects in statistical measurements.

The task **Creation of joint micro data** is related to the GSBPM sub-process 5.1 "Integrate data". According to the definition in UN/ECE (2013), *this sub-process integrates data from one or more sources. It is where the results of sub-processes in the "Collect" phase are combined.* Matching / record linkage routines are explicitly mentioned in this process.

Multisource estimation at aggregated level is related to the sub-processes 5.1 "Integrate data", 5.6 "Calculate weights", 5.7 "Calculate aggregates". As far as sub-process 5.1 is concerned, it is important to remark that according to the definition given in GSBPM, "the result is a set of linked data", given that in GSBPM it is explicitly mentioned "combining data from multiple sources, as part of the creation of integrated statistics such as national accounts", we interpret data in its more general meaning, that is micro-data and data at aggregated level. The sub-process 'calculate weights' is related to the cases when administrative data are used jointly with survey data.

Finally a table cross-classifying usages and statistical tasks is depicted in the following Table 1.

In the table, usages are detailed according to the specific usages described in Deliverable 1. White cells indicates when the statistical task is not generally required for that specific usage.

Table 1. Relationships among usages (specific and not) and statistical tasks.

		Usages													
		Direct usages				**Indirect**									
		Direct tabulation		**Substitution and Supplementa-tion**		**Creation and maintenance of registers and survey frames**		**Editing and imputation**			**Indirect estimation**			**Data validation, confrontation**	
Specific usages (see below)		***su1***	***su2***	***su3***	***su4***	***su5***	***su6***	***su7***	***su8***	***su9***	***su10***	***su11***	***su12***	***su13***	***su14***
Statistical tasks	**Data editing and imputation**														
	Unit harmonisation														
	Measurement harmonisation														
	Creation of joint data														
	Multisource estimation at aggregated level														

In order to improve the readability of Table 1, the specific usages (as described in Deliverable 1) are listed.

Direct Usage

Direct Tabulation

Specific usages

su1. Exploiting only one administrative data source

su2. Exploiting multiple administrative data sources

Substitution and Supplementation for Direct Collection (Replacement for data collection)

Specific usages

su3. Split-population approach

su4. Split-data approach

Indirect usage:
Creation and maintenance of registers and survey frames

Specific usages

su5. Identification of frame units and their connections to population elements

su6. Identification of classification and auxiliary variables

Editing and imputation

Specific usages

su7. Construction of edit rules

su8. Construction of models to find errors in data

su9. Auxiliary data to construct imputation models

Indirect estimation

Specific usages

su10. Creation of population benchmarks to be used for calibration

su11. Use administrative data in a predictive setting

su12. Estimation where administrative and statistical data are used on an equal footing

Data validation/confrontation

Specific usages

su13. Validation of survey estimates and/or other administrative data sources

su14. Address quality issues

References

Deliverable 1. Identification of the main types of usages of administrative sources.

Groen J.A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures, *Journal of Official Statistics vol. 28, n. 2, pp.* 173-198

Memobust (2014). Statistica Data Editing , in *Memobust Handbook on Methodology of Modern Business Statistics.* https://ec.europa.eu/eurostat/cros/content/memobust_en

Memobust (2014a). Editing administrative data, in *Memobust Handbook on Methodology of Modern Business Statistics.* https://ec.europa.eu/eurostat/cros/content/memobust_en

UN/ECE (2011). Applying the Generic Statistical Business Process Model to business register maintenance. UN/ECE Conference of European Statisticians, Group of Experts on Business Registers, Twelfth session, Paris, 14-15 September 2011.

UN/ECE (2013), Generic Statistical Business Process Model. Version 5.0 – December 2013. The United Nations Economic Commission for Europe (UN/ECE). See: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0.

UN/ECE (2014). Measuring population and housing. Practices of UNECE Countries in the 2010 round of censuses. United Nations Economic Commission for Europe.