Estimation methods for the integration of administrative sources

Task 1: Identification of the main types of usages of administrative sources

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 D1**
Date of first version:	30.08.2016
Version:	V3
Date of updated version:	19.12.2016.
Written by:	Ton de Waal, Sander Scholtus, Arnout van Delden, Marco Di Zio, Nicoletta Cibella, Tiziana Tuoto, Mauro Scanu, with consultations with Li-Chun Zhang
Sogeti Luxembourg S.A.	Laurent Jacquet (project manager)
	Sanja Vujackov

Table of contents:

Introduction

Administrative data can be used for different purposes in the statistical production system. This document identifies the main usages of administrative sources in the practice of National Statistical Institutes (NSIs) and its aim is to provide the reference for connecting the statistical integration methods to the statistical production problems involving the use of administrative data.

From the outset, some key words need to be clarified: administrative data, statistical data and integration.

For the definition of administrative data, it was used the one proposed by UN/ECE (2011): administrative data is "*data holding containing information which is not primarily collected for statistical purposes*". It is further noted that in the wide sense this definition aims to cover any potentially useful 'non-statistical' data that are "not primarily collected for statistical purposes". Examples include various commercial or financial transaction data, sensor-generated data, social media data, etc.

We remark that, even though this definition includes big data, due to their peculiarity, they are not considered in the rest of the paper. Data from administrative systems at various government bodies are usually the most readily accessible with regard to the necessary legal, operational and technical considerations. Depending on the context, it is possible to use the term "administrative systems data" where one wishes to convey the narrower interpretation.

The term statistical data is then used in this document to represent censuses and survey data, which are directly collected for statistical purposes. Notice that historically speaking statistics produced from administrative systems data were among the first government statistics (Nordbotten, 2012). Indeed, whenever administrative data are used without any direct involvement of census and survey data, one must evaluate the administrative data statistically, hence treat them as statistical data. The distinction between "statistical" and "administrative" data as used in this document, therefore, refers above all to the primary and second nature of their uses.

For statistical integration we refer to any process where more than one data source is used and statistical methods are needed for the use of those sources. Statistics produced by statistical integration are sometimes referred to as multi-source statistics.

Even though research covered many different research papers and guidelines, this document is mainly focused on Brackstone (1987), Statistics Canada Quality Guidelines (2009) and MIAD (2014).

Related works

Although there is still a lively discussion on the use of administrative data in the production of official statistics, studies on this problem have a quite long history. In Brackstone (1987), written thirty years ago, there is an interesting discussion about problems and perspectives on the use of administrative data, and therein can be found a list of possible classes of usages:

1. ***Direct tabulation***

2. ***Indirect estimation***
   Administrative data comprise one of the inputs into an estimation process.

3. ***Survey frames***
   Administrative data used to create, supplement or update frames to be used for census and surveys.

4. ***Survey evaluation***
   Administrative data used for checking, validating and evaluating survey data.

Those usages and the relevant descriptions are still valid, but the list is influenced by the view of the time the paper was written and of the context in which it is written. More recently, other documents about administrative data illustrate their possible usages. It is interesting to report the one in the Statistics Canada Quality Guidelines (2009), the same organisation to which Brackstone belonged but about twenty years later:

(i) use for **survey frames**, directly as the frame or to supplement/updated an existing frame;
(ii) **replacement** of data collection (e.g. use of taxation data for small businesses in lieu of seeking survey data for them);
(iii) use in **editing and imputation**;
(iv) **direct tabulation**;
(v) **indirect** use in estimation;
(vi) **survey evaluation**, including data confrontation.

We notice that, compared to the previous, there are two more elements mentioned in this classification: (ii) replacement of data collection and (iii) use in editing and imputation. Both elements were actually included in the Brackstone's classification, and use (ii) was even explicitly mentioned under the "indirect estimation".

The need of stating them as separate categories testifies the increasing weight given to administrative data in the production of official statistics. In particular, the replacement of data collection is characterised by an extensive use of administrative data.

In the European context, a description of usages can be found in the European project 'Admin Data'. This project is composed of European Countries and it is aimed at exploring the possibilities of using the administrative data for business statistics[1].

The list of usages is:

1. in statistics production as a replacement for primary and/or complementary data to other sources;
2. as a sample framework and source of auxiliary information in sample design;
3. as a source of additional variables to be used for estimates;
4. as auxiliary information to support processing of primary data as for example: data editing, imputation, calibration of estimates;
5. as input for statistical registers based system.

[1] For more information see http://ec.europa.eu/eurostat/cros/.

In this list, the weight given to the administrative data for the production of statistics is even stronger. It can be actually observed that the last bullet (5), which most exploits the admin data, is in fact not excluded from the classification given by Statistics Canada papers. However, countries participating in above-mentioned project wanted to explicitly mention this as a separate point, perhaps partly to emphasize the distinction between a statistical system and any specific statistics.

List of usages

In order to choose one of the lists available in the observed literature, which are indeed almost the same, it helps to keep in mind the purpose of introducing such classification for this project.

In a so-called fit-to-use approach, the aim is to connect the practical usages of administrative data in NSIs to statistical integration methods that are potentially useful for that practice. What is the most important is not really the number of categories, or even to some extent under which category each usage is classified, but the fact that the different main usages are covered.

With this idea in mind, we have adopted the document produced within the MIAD project (Methodologies for an Integrated Use of Administrative Data in the Statistical Process) as the base of the list of usages (MIAD, 2014). While it is consistent with all the others, this document details more in depth how administrative data are used for producing statistics in NSIs. However, for the purpose of this project, we have made some changes to it, which will be explained further in this document, and added relevant comments at different places.

Although not essential for our purpose, it is interesting to introduce firstly a distinction which is whether administrative data enter the production of statistics in a direct or an indirect way.

Following the definition given in MIAD, a "***direct usage*** *will be defined as situations where there is an immediate link between the ADS [i.e. administrative data sources] and statistical output. The ADS may undergo various transformations, such as converting administrative units to statistical units (e.g. profiling businesses obtained from tax registers), or deriving statistical output variables from the unit's attributes, but in essence output is primarily sourced from the ADS itself.*", we notice that this covers also the case where relevant statistical data are available but are only used to help defining the processing of the ADS.

Indirect usages of administrative data are those "*situations where the ADS plays a supporting role in the creation of statistical output sourced primarily from either a survey or another ADS.*

Examples include the use of ADS to create a survey frame, or as population benchmarks used in weighting sample data." We notice that the term "supporting" does not mean that the ADS is 'helpful but unnecessary' in such cases. For example, "a survey frame" is not only necessary but is often of critical importance to the output. Neither does the term "supporting" imply the ADS otherwise provides only 'auxiliary data'. For example, population size estimation based on ADS and coverage surveys would not have been possible at all without the ADS enumerations. Thus, the term "supporting" only means that the ADS does not suffice on its own; and one should keep in mind that neither would the statistical data necessarily have sufficed on its own.

The following part of the document describes briefly the usages and the main integration problems that are encountered in practice.

I Direct Usage

Direct usage covers Direct tabulation and Substitution and Supplementation for Direct Collection.

1. Direct Tabulation

Description: The case where administrative data are used to produce statistics without resorting to any statistical data.

Specific usages:

su1. Exploiting only one administrative data source

Typical examples of statistics produced according to this framework are 'International trade statistics', statistics on vital events like numbers of births and deaths, crime rates, etc. Note that when there is one administrative data source, data integration is not necessary.

su2. Exploiting multiple administrative data sources.

Important cases are the register-based census-like statistics in a number of European countries, UN/ECE (2014).

We notice that the use of the base registers, i.e. the Population Register, the Business Register and the Immobility Register (including address, property and land), is often necessary in the case of "only one administrative data source", in order to delineate the target population. The base registers themselves of course originate from administrative systems data. However, it is possible to consider the base registers as part of the statistical infrastructure, and leave them out of the specific context here.

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

Description: The case when 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.

Specific usages:

su3. Split-population approach

In this model the statistical population is split into two or more parts for data collection purposes. Data from administrative sources are used for units where these data are of sufficient quality, and statistical sources are used for the remainder of the units.

A typical scenario for a business survey is that data for relatively small businesses with simple structures are taken or derived from tax returns, whereas surveys are used to collect data from the key units (usually those that are largest and/or have the most complex structures). For the section of the population for which tax data are used, the statistical and administrative units are likely to be identical, or very similar, and the impact of the difference between statistical concepts and classifications and their administrative counterparts is likely to be minimal, or at least can be easily remedied, for instance by rule-based processing[2]. An example can be found in Delden et al. (2016) where the turnover is derived from tax returns by computing (for part of the domains) a correction factor that needs to be updated regularly.

su4. Split-data approach

In this approach, a population of statistical units and a data requirement are identified, for example the population could be all persons living in a particular country, and the data requirement could be the usual set of variables required for a population census. Instead of providing all of the variables for part of the population, as in the split population model above, under the split data approach, administrative sources are used to provide some of the variables for all of the population units[3]. An example is in Luzi et al. (2014), where some key variables for structural business statistics are estimated mainly by using the administrative data.

We notice that Brackstone (1987) in fact characterizes the split-population and split-data approach under "indirect estimation". The key point is that here the data from different sources are supplementary of each other literally, cf. the indirect usages below where the data may be 'overlapping' of each other.

[2] See UN/ECE, 2011, Chapter 8.
[3] See UN/ECE, 2011, Chapter 8.

II Indirect Usage

Indirect usage covers:

- Creation and maintenance of registers and survey frames
- Editing and imputation
- Indirect estimation
- Data validation/confrontation

1. Creation and maintenance of registers and survey frames

Description: Administrative data are used for creating and maintaining registers and survey frames.

Specific usages:

Sampling frames require the:

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

su6. Identification of classification and auxiliary variables (e.g. NACE or size of business).

We notice the subtle distinction between classification and auxiliary variables in a frame. For example, in business surveys, both the NACE[4] and some measure-of-size variables are used in the design of stratification and inclusion probabilities.

The number of employees is the most common measure-of-size variable in practice. Sometimes, past turnover in taxation data and other administrative data are also used, which are usually referred to as frame auxiliary data. The point is that while frame auxiliary data is important for the efficiency of the sampling design, the NACE is fundamental because a unit that is wrongly classified to be out of the target population will have no chance of being included in the sample, hence violating the principle of probability sampling.

Notice also that we have deleted the usage "Construction sampling designs" in the MIAD list, because the uses of measure-of-size variables and identifying variables are already included as part of the frame discussed above. It also seems unnecessary to separate a frame from its use for sampling design.

2. Editing and imputation

Description: Administrative data are used to check and impute survey/administrative data.

Specific usages:

su7. Construction of edit rules

For instance a ratio edit between the observed number of employees in a survey and the number of employees in the Business Register; an application about the construction of tolerance ranges for the *population* domain is ONS (2012).

[4] The NACE is referred to as a domain classification variable that is considered to be of critical importance in business statistics.

su8. Construction of models to find errors in data

For instance, balance sheets and taxation data may be used in a model for predicting errors in the observed survey data on Business investments (Di Zio et al., 2015). The predicted error is used for selective editing, that is to prioritise observations to be carefully revised.

Another application can be found in Guarnera et al. (2016) to estimate the labour cost with the Financial Statements from the Chamber of Commerce and data based on social security information. Data from Financial Statements are taken as reference data, however there are specific situations where those values are not coherent with the statistical definition. This is the case of costs for workers like *agency workers* and *external workers* (for example *project workers*) that should be excluded from personnel costs according to Structural Business Statistics (they should be included in the *intermediate costs*), but that are usually not distinguished from the costs for employees in the company accounts. A latent model is used to correct the values reported in financial statements by using data from social security.

su9. Auxiliary data to construct imputation models

For instance, balance sheets and taxation data may be used in a model for imputing missing data, or, in so-called calendarization, a quarterly ADS value may be broken down to impute monthly survey data that are missing.

3. Indirect estimation

Description: Administrative data are one of the inputs of the estimation process.

Specific usages

su10. Creation of population benchmarks to be used for calibration

su11. Use administrative data in a predictive setting

Variables observed in administrative data for all the units in the population are used to predict values for variables gathered only in a sample for the non-sampled units (Luzi et al. 2014), including small area estimation.

Using the early available administrative data to predict later more reliable estimates, where the latter requires time for the data to 'mature' (e.g. Zhang and Pritchard, 2013).

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

For instance,

- population size estimation based on ADS enumerations and coverage surveys;
- estimation based on latent class models where ADS and statistical variables are both subjected to measurement errors;
- benchmarked mass imputation for census-like output tables or small area estimation;
- reconciliation of multiple time series from different sources and with different frequencies;
- balancing of supply-and-use tables for GDP; etc.

We notice that survey weighting and prediction modelling are traditional approaches of indirect estimation. However, estimation where administrative and statistical data are used on an equal footing covers many emerging important and potentially highly rewarding uses of administrative data for statistical purposes. For example, it may enable one to produce census-like statistics, without the traditional census, on a much more timely basis and with greatly reduced costs. Similarly, it can greatly extend the scope of small area estimation compared to what is achievable under prediction modelling.

4. Data validation/confrontation

Description: Administrative data are used to validate data.

Specific usages:

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

Survey estimates are compared to specific aggregates in order to verify whether there is an anomalous behaviour. Another example is where suspicious micro-data from a survey values are compared with (time series of) administrative data to manually check the plausibility of the micro data.

su14. Address quality issues

Many quality aspects are evaluated by using administrative data, e.g. a recent paper discusses the problem of evaluating of classification errors in the NACE code (Burger et al., 2015).

A more extensive description of some applications concerning usages will be given in the document produced by Task 4.

Usages in terms of data configurations

In the ESSnet on quality of multisource statistics (Komuso) a classification of six so-called basic data configurations has been proposed. These basic data configurations seem the most important ones in practice for structuring the work on quality measurement of multisource statistics. In the current project we will use a version of the Komuso classification that has been adapted for structuring estimation methods for multisource statistics.

In this adapted version eight basic data configurations are distinguished. We will discuss the connection of these eight basic data configurations with the usages of estimation methods.

Many practical situations can be built on these basic data configurations. Although not exhaustive, it covers most of the practical situations arsing in estimation with multisource data. Supporting activities such as the creation and maintenance of registers and frames, editing and imputation, and data validation/confrontation are not considered explicitly in these classifications.

The other possible usages of administrative data aforementioned such as Direct Tabulation, Substitution and Supplementation for Direct Collection, and Indirect estimation, are considered in the original Komuso classification as well as in its adapted version. For a more extensive discussion of these basic configurations, their problems and solution methods see De Waal, Van Delden and Scholtus (forthcoming).

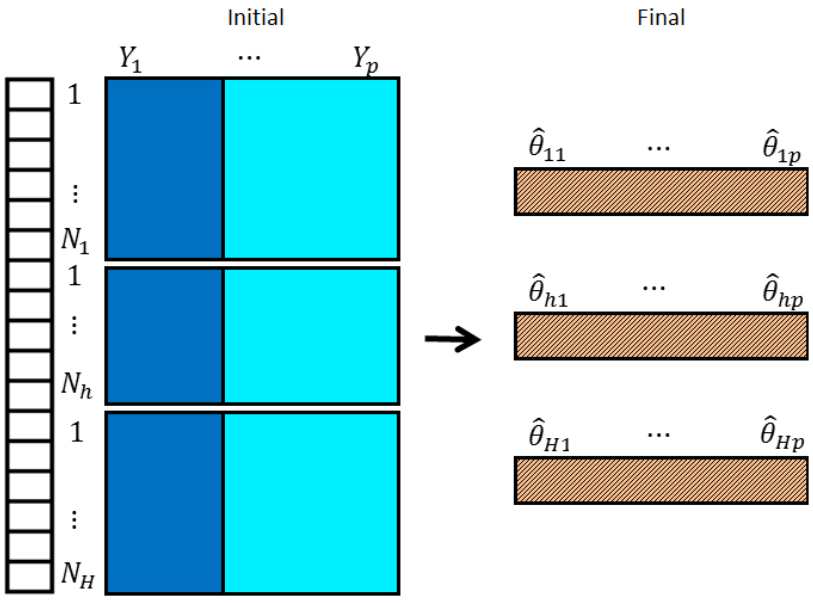
Basic data configuration 1

The first and most basic data configuration concerns multiple cross-sectional data sources covering the target population where the different data sets contain different target variables. We refer to this as the "split-variable" case. An example of this is the population census where in some countries most variables are collected through a census survey data and some variables are obtained from administrative sources. The administrative sources will have to be linked to a population register, which requires linkage techniques. This linkage step is also needed in the basic data configurations 4 and5. After linkage, provided that the data are error-free, the data can simply be "added" to produce output statistics. This basic data configuration is a stepping stone to other situations. We will call this Basic data configuration 1.

Basic data configuration 1 is illustrated in Figure 1. Concerning the illustrations in this document note that:

1) The rectangle of white blocks to the left represents the population frame;
2) Different blue colours represent different input data sources
3) Orange/brownish colours represent derived output statistics
4) Shaded blocks represent macro data, bright blocks represent micro data.

Figure 1. Combining non-overlapping microdata sources without coverage problems, "split-variable" case



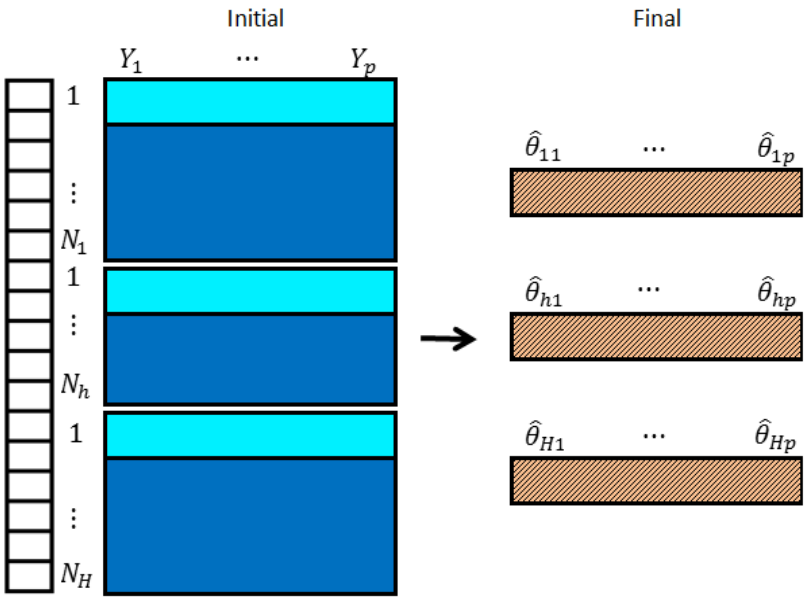
When all the sources in the split-variable case are administrative sources, then this data configuration is an example of the "direct tabulation" situation. When it concerns a combination of

administrative and survey data it concerns "substitution and supplementation for direct collection". However, in this case it concerns a census survey, where all units have a weight of one. In terms of the required methods, the latter case will resemble the "direct tabulation" case.

Basic data configuration 2

The second basic data configuration also concerns multiple cross-sectional data sources covering the target population but in this case the different data sets contain different units, i.e. the "split-population" case. This situation can be more complicated than basic configuration 1, because there might be small differences in the concepts of the variables. That implies that the variables in both sources need to be harmonised. Provided the data are in an ideal error-free state, the different datasets, or data sources, are ***complementary*** to each other in this case, and likewise to basic data configuration 1 they can simply be "added" to each other in order to produce output statistics.

Figure 2. Combining non-overlapping microdata sources without coverage problems

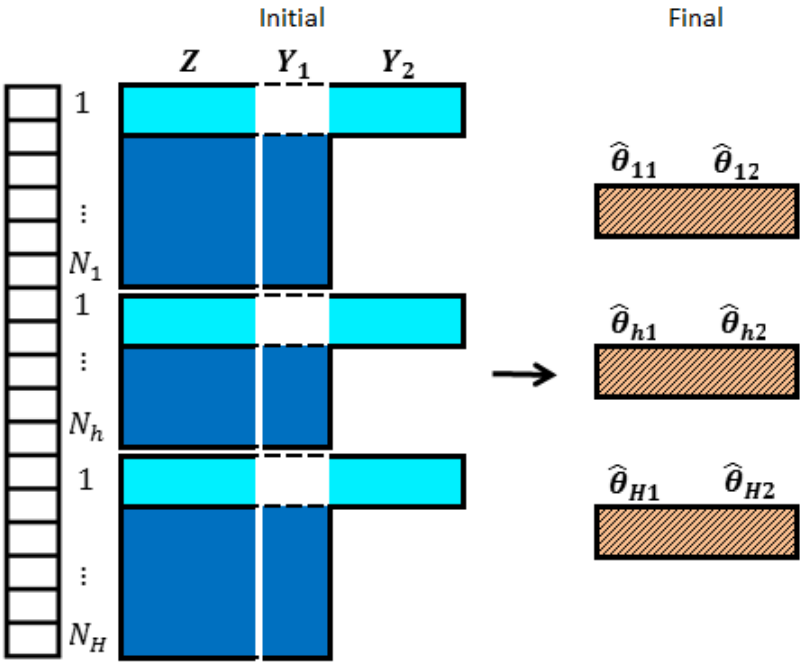


As with data configuration 1, when all the sources in the split-variable case are administrative sources, then this data configuration conforms to the usage "*direct tabulation*". When the *split-population* situation concerns a combination of administrative and survey data it complies with the usage "*substitution and supplementation for direct collection*".

Basic data configuration 3

A slightly different situation occurs when we have non-overlapping units, as in basic data configuration 2, we have a number of overlapping variables, but also some target variables that are available in only one of the sources. We call this Basic data configuration 3 which is illustrated in Figure 3. In this figure, variables $\boldsymbol{Z}$, where $\boldsymbol{Z} = (Z_1, \ldots, Z_k)'$, are the common (background) variables that are used to match the data sources. In addition, data source 1 contains variables $\boldsymbol{Y}_1$ and data source 2 $\boldsymbol{Y}_2$. We still like to join also the non-overlapping variables to the other units at micro-level. We could use statistical matching techniques, where overlapping variables are used to find similar units in the data sources (see, e.g., D'Orazio, Di Zio and Scanu 2006 for more on statistical matching).

Figure 3. Combining non-overlapping micro data sources with part of the variables is in a single source, without coverage problems



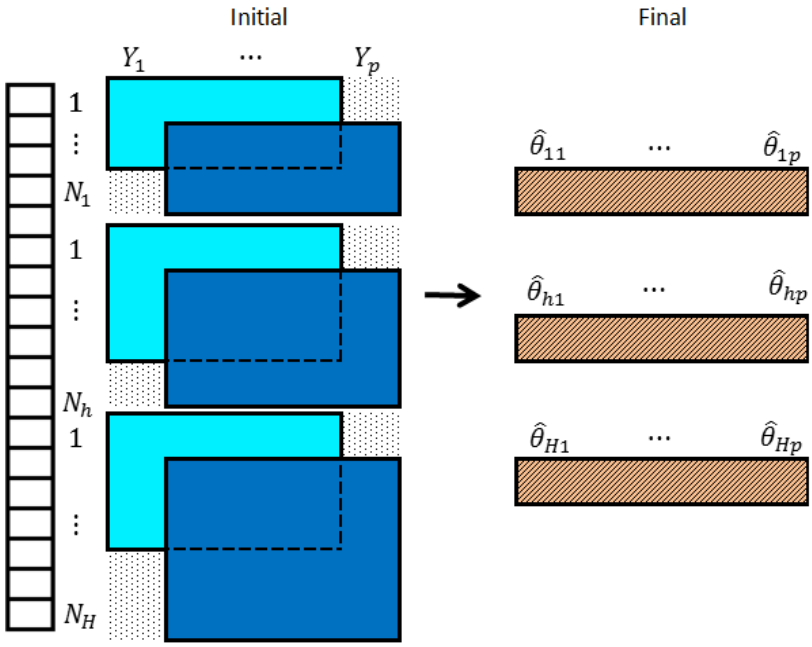
Note that a small modification of basic configuration 3 with the situation where the variables in one source (dark blue) is subset of the variable in the other source (light blue). In that case, the non-overlapping variables may be calibrated on outcomes of the overlapping variables; see, e.g., Särndal et al. (1992).

Basic configuration 3 might sometimes concern a situation with only administrative data sources ("direct tabulation") , where the difference sources do not have a complete overlap in the variables. More likely however is that it concerns a combination of administrative with survey data (the usage type "substitution and supplementation for direct collection"). For instance one may have data from business administrations for a large set of units, and a census survey sample for a limited set of units, and the variables in both sources partially overlap.

Basic data configuration 4

Basic data configuration 4 is characterised by a deviation from basic data configuration 2, by which there exists ***overlap*** between the different data sources. The overlap can concern the units, the measurements, or both. Basic data configuration 4 is illustrated in Figure 4.

Figure 4. Combining overlapping micro-data sources without coverage problems



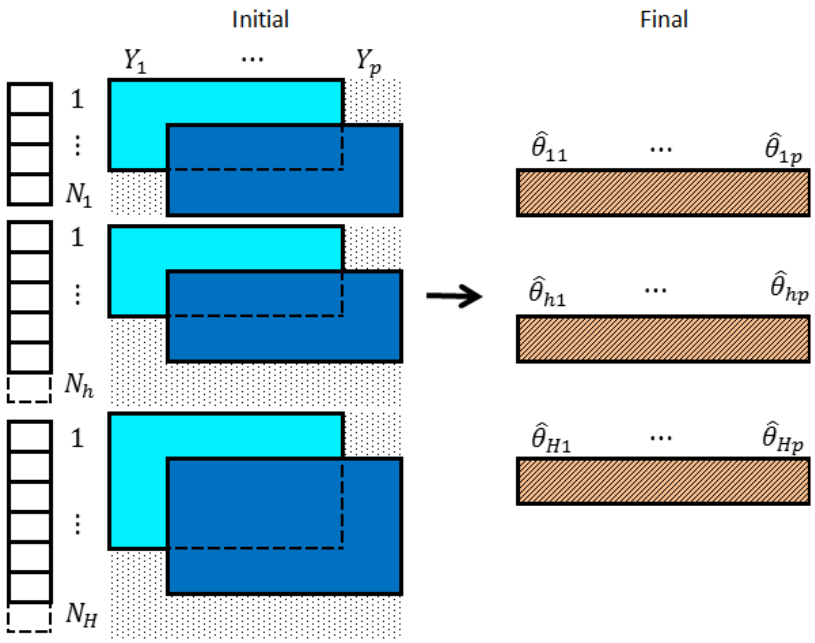
When the same phenomenon is observed for the same units in multiple data sources, one can utilize the multiple observations to identify and correct residual errors. Differences between the responses of a unit are caused by measurement error in one (or more) of the observed values.

The actual usage of administrative data in this data configuration depends on sources and the estimation methods. When all data sources are administrative, then "direct tabulation" applies. When data sources are a combination of administrative and survey sampling data and micro-integration is used, then the usage type "substitution and supplementation for direct collection" applies. When data sources are a combination of administrative and survey sampling data and measurement errors in both sources are modeled then *indirect estimation* applies.

Basic data configuration 5

Basic data configuration 5 is characterised by a further deviation from basic data configuration 4, by which the combined data entail ***under coverage*** of the target population in addition, even when the data are in an ideal error-free state. Basic data configuration 5 is illustrated in Figure 5.

Figure 5. Combining overlapping micro-data sources with under coverage



A possible example of Basic data configuration 4 is a population census followed by a post-enumeration survey (PES). Both the census and the PES entail under-count of the target population. The binary measurement of being enumerated is the overlap between the two.

Besides under coverage also over coverage, i.e., units not belonging to the target population are in the data sources .Record linkage techniques may be used to remove duplicated units, while model based approaches may be useful to classify whether units belong to the target population.

Like basic data configuration 4, basic data configuration 5 concerns either usage "direct estimation", "substitution and supplementation for direct collection" or "indirect estimation".

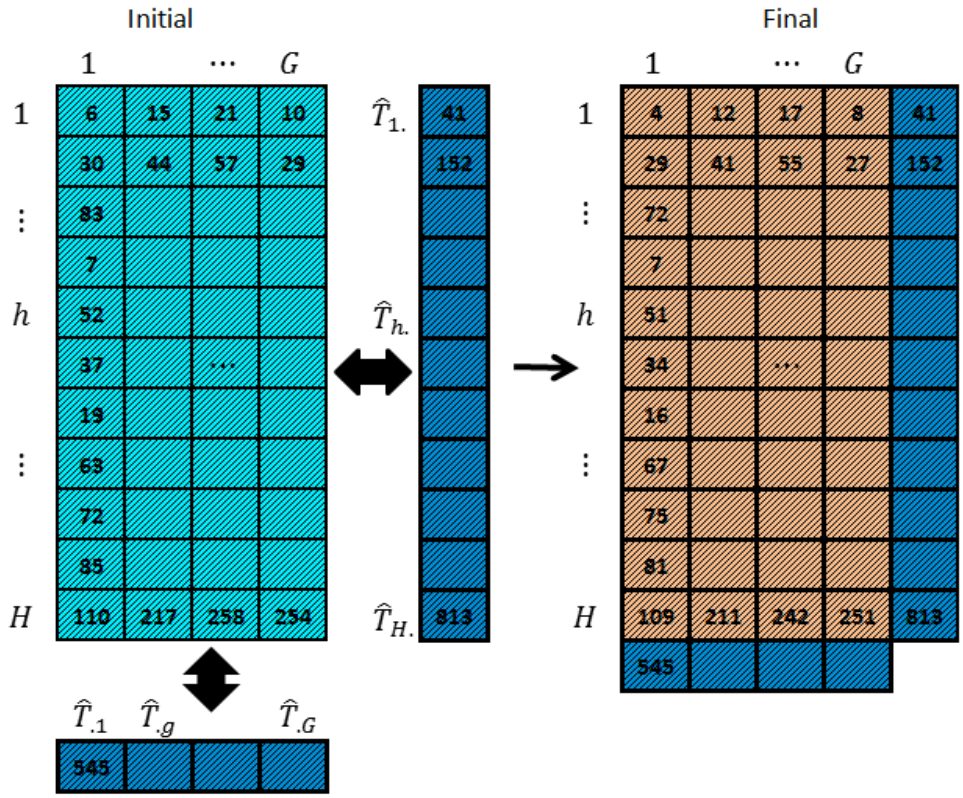
Basic data configuration 6

Basic data configuration 6 is the complete macro-data counterpart of Basic data configuration 4: in basic data configuration 6 only aggregated data overlap with each other and need to be reconciled. Basic data configuration 6 is illustrated in Figure 6.

An example of Basic data configuration 6 is provided by the National Accounts, where aggregated data from many different sources need to be reconciled with each other subject to both equality and inequality constraints. To reconcile only aggregated data macro-integration can be applied.

Data configuration 6 is a typical example of the indirect usage of administrative data, e.g. for balancing supply and use tables.

Figure 6. Combining macro-data sources



Basic data configuration 7

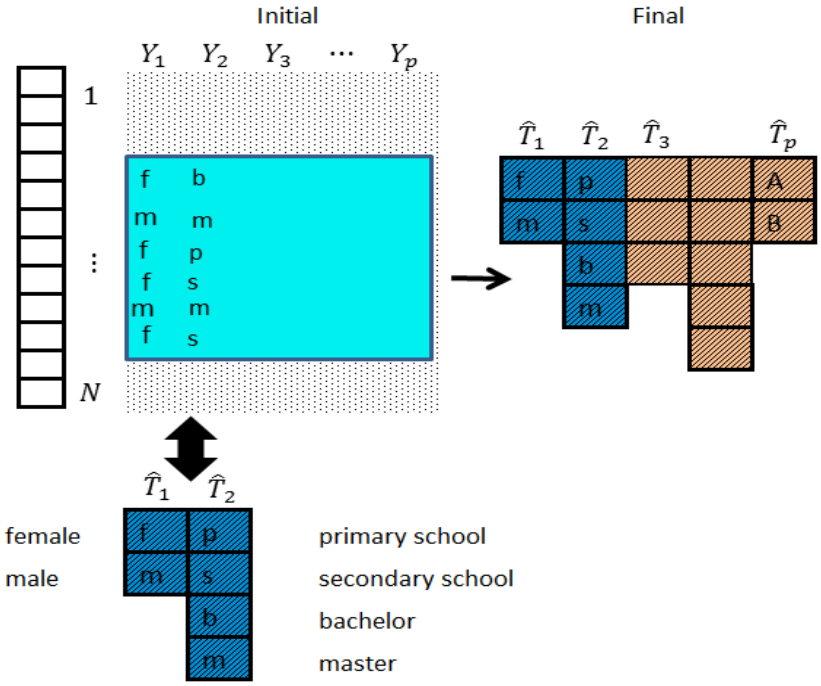
We will now look at basic data configurations where we have cross-sectional macro-data, possibly in combination with micro-data. Basic data configuration 6 is characterised by a variation of Basic data configuration 3, by which ***aggregated data*** are available besides micro data.

There is still overlap between the sources, from which there arises the need to reconcile the statistics at some aggregated level. An important special case occurs when the aggregated data are estimates themselves. Otherwise, the conciliation can be achieved by means of calibration which is a standard approach in survey sampling.

Basic data configuration 6 is illustrated in Figure 7.

Like data configuration 6, data configuration 7 is another example of the indirect usage of administrative (e.g. benchmarked imputation for census-like output tables, see task 1).

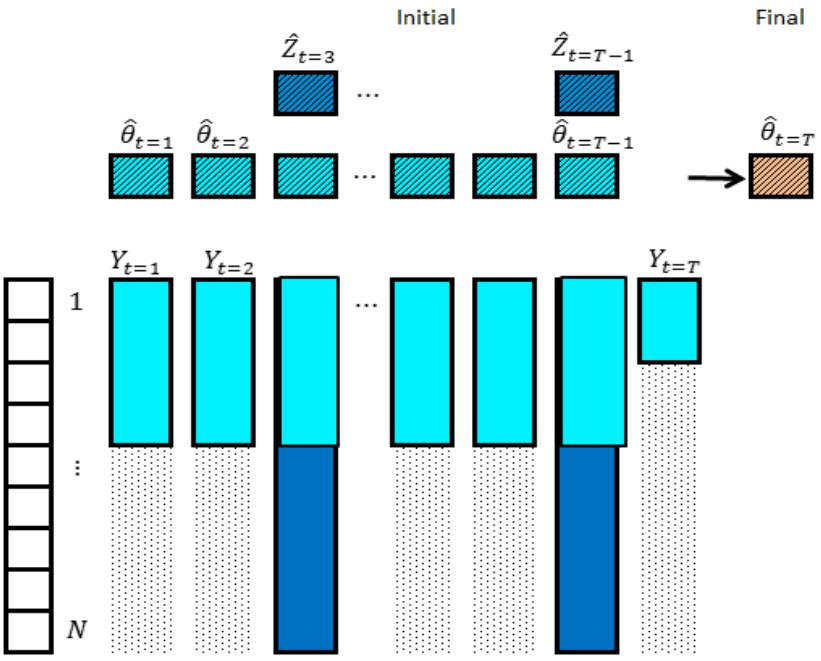
Figure 7. Combining a micro-data source with a macro-data source



Basic data configuration 8

Finally, longitudinal data are introduced in Basic data configuration 8. We limit ourselves to the issue of reconciling time series of different frequencies and qualities, as illustrated in Figure 8.

Figure 8. Combining longitudinal data sources



An example of Basic data configuration 8 is when turnover data of enterprises are available on a monthly basis from a survey and on a quarterly basis from the Tax Office. The monthly survey data and the quarterly administrative data may be combined to produce a new, benchmarked time series. We want to develop methods for this.

Like data configurations 6 and 7, data configuration 8 concerns indirect usage of administrative (e.g. reconciliation of multiple time series from different sources and with different frequencies, see task 1).

Integration problems

In general, all the cases require one to deal with:

a. the problem of univalency (i.e. the same number for the same phenomenon);
b. a harmonisation of different data sources;
c. a procedure for linking or matching data sources.

Since administrative data are collected with purposes generally different than that of our objective, first needed is a process of evaluation of proximity or univalency[5] both in terms of statistical units and measurements, i.e., the statistical units are essentially the same in the different data sources, and the variables measure the same concept. Consequently, when needed, a process of harmonisation or alignment must be performed. This means that the objects observed in the administrative data sources must be transformed into statistical units related to the target population, and that the attributes characterising the objects must be transformed into measurements related to the objective of our investigation. An example is when the variable concerned with 'age' is observed in two data sources but with different classes. A task for obtaining a unique variable 'age' is needed. This apparently deterministic task involves statistical estimation methods whenever it is not known a well-defined mapping. Another example is in D'Orazio et al. (2016), where an integration of the Italian Household Budget Survey (HBS) with the Banca d'Italia Survey on Household Income and Wealth (SHIW) is described. An important variable for this problem could be the 'head of household'. Although both of them gather information on it, the definition is different, the HBS takes the head of the household to be whoever is already registered in the public archives, while SHIW assumes that the head of the household is the person responsible for the household finances. These two definitions cannot be harmonized through the available information.

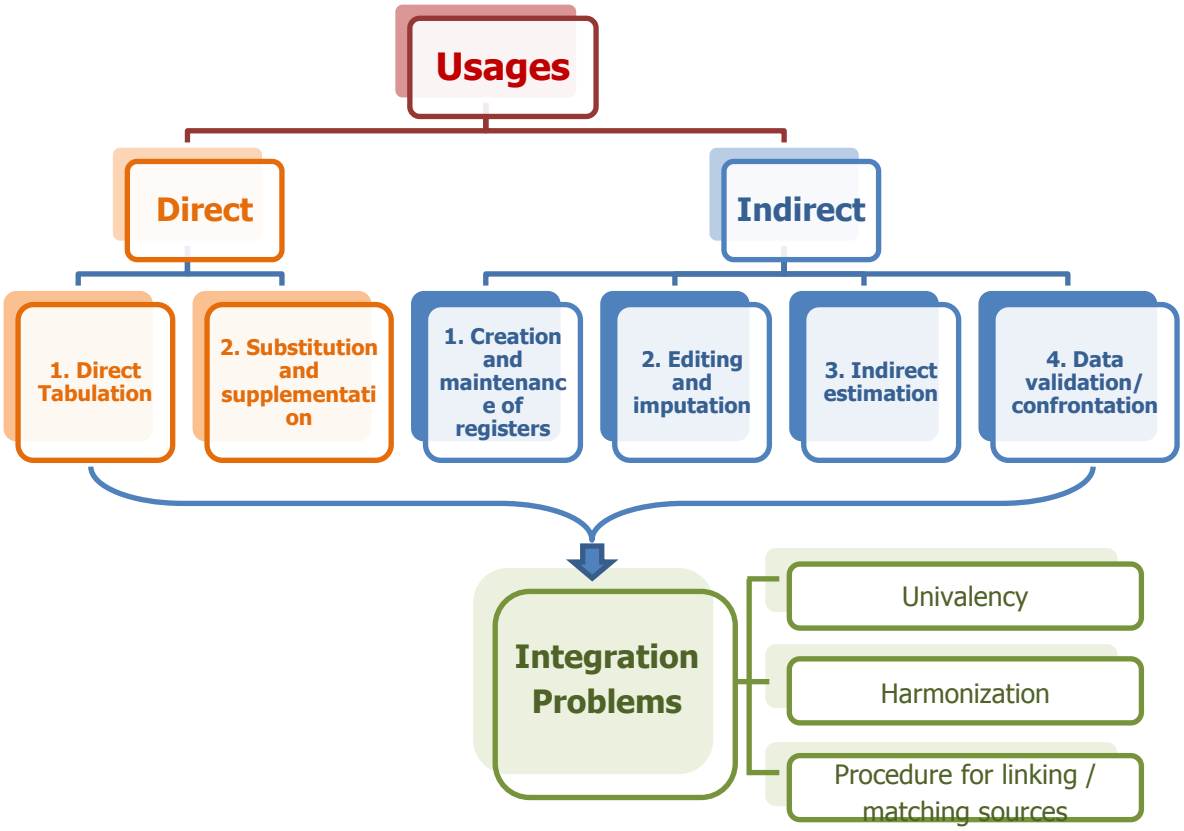
[5] The concepts of coherence and comparability described in EES QAF (2015) include the concept of univalency, but they are wider. For instance they may refer to the consistency of different variables, e.g., turnover and costs , while univalency is referred to the consistency of the same variable, e.g., turnover, measured in two or more data sources.

A further important step in order to use administrative data in an integrated framework is concerned with the linking or matching of the different data sources. Record linkage is referred to the linkage at micro level, that is when we need to identify the same unit in two data sets. Statistical matching is an integration at micro and macro level, and it is applied in case two (or more) data sets do not contain the same units.

For all these tasks, it is usually necessary to resort to statistical methods of estimation. Also, it is certainly always necessary to use the statistical methods to evaluate uncertainty related to the attained conclusions. For instance, in probabilistic record linkage the evaluation of the errors in the matching procedure and their impact on the target estimate (e.g., a population total) would be desirable, see Di Consiglio and Tuoto (2015).

Figure 1 synthetically represents the list of usages and integration problems.

Figure 1. List of usages of administrative data sources and integration problems



The two-phase life-cycle model for integrated statistical micro-data of Zhang (2012) describes in more details which are the tasks characterising the previous integration problems. Although that paper focuses on micro-data, many of the ideas expressed by the Author may be used in the context of macro-data integration as well, especially for the aspects concerning 'measurements'. A graphical description of the two-phase life cycle is provided in Appendix A.

References

Brackstone, G.J. (1987) Issues in the use of administrative records for statistical purposes, *Survey methodology,* vol. 13, n. 1, pp 29-43.

Burger, J., Scholtus, S., Van Delden A. (2015) Sensitivity of Mixed-Source Statistics to Classification Errors. *Journal of Official Statistics.* Vol. 31, n. 3, pp. 489-506.

Delden, A. van, Pannekoek, J., Banning, R. and de Boer A. (2016). Analysing correspondence between administrative and survey data. *Statistical Journal of the IAOS* (in press).

De Waal, T., A. Van Delden and S. Scholtus (forthcoming), *Multisource Statistics: Basic Situations and Methods*. (tentative title)

Di Consiglio, L., Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics.* Vol. 31, n. 3, pp. 415-419.

D'Orazio, M., Di Zio, M., Scanu, M. (2016). *Statistical matching: theory and practice*. Wiley & Sons,. Chichester.

Di Zio, M., Guarnera, U., Iommi, M., Regano, M. (2015). Selective editing of business investments by using administrative data as auxiliary information. *Unece, Worksession on Statistical Data editing, Budapest, Hungary, 14-16 September 2015.*

ESS QAF 2015. Quality Assurance Framework of the European Statistical System. v1.2 http://ec.europa.eu/eurostat/web/quality

Guarnera, U., Pacini, S., Varriale, R., (2016). A latent class model to estimate labour cost from multisource data, *proceedings of European Conference on Quality in Official Statistics (Q2016) Madrid, 31 May-3 June 2016.*

Luzi, O., Guarnera, U., Righi, P. (2014). The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data. *European Conference on Quality in Official Statistics (Q2014). Vienna, 3-5 June.*

MIAD (2014). Activity A: Mapping and Overview. Usage of Administrative Data Sources for Statistical Purposes. https://ec.europa.eu/eurostat/cros/content/miad-methodologies-integrated-use-administrative-data-statistical-process_en.

Nordbotten, S. (2012). The Use of Administrative Data in Official Statistics - Past, Present, and Future - With Special Reference to the Nordic Countries. *Official Statistics in Honour of Daniel Thorburn*, 205–223.

ONS (2012). Using administrative data to set plausibility ranges for population estimates. Research report.

Särndal, C.-E., B. Swensson, and J. H. Wretman, 1992, *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Statistics Canada Quality Guidelines (2009). *Use of Administrative Data*, http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm.

UN/ECE (2011). *Using Administrative and Secondary Sources for Official Statistics.* A Handbook of Principles and Practices.

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

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

Zhang, L.-C. and Pritchard, A. (2013). Short-term turnover statistics based on VAT and Monthly Business Survey data sources. *ENBES workshop 2013, Nuremberg*.

Appendix A

The two-phase life cycle of integrated statistical micro data is summarised in Figure A1. It describes the various states of data (rectangles) referring to 'measurements' (variables) and 'representation' (objects), and the potential errors (ovals) that can occur in that phase. The first phase is concerned with the life cycle of a single source, the second phase refers to the phase of integration of data sources.

Figure A1. The two-phase life cycle of integrated statistical micro data (Zhang, 2012).