Guidelines on the use of estimation methods for integrating administrative sources

Authors:

Wilfried GROSSMANN

Mauro MASSELLI

Manlio CALZARONI

Carlos DIAS

Patrícia XUFRE

Marco DI ZIO

Knut UTVIK

Rev 2.1
2019

Table of contents

Figures

Acronyms

BR	Business Register
EBLUP	Empirical Best Linear Unbiased Predictor
EDI	Electronic Data Interchange
EEC	European Economic Community
EOTE	Employment Outcomes of Tertiary Education
EQLS	European Quality of Life Survey
ESS	European Statistical System
ESSnet	Collaborative ESS networks
ESSnet ISAD	ESSnet on Integration of Survey and Administrative Data
ESS.VIP ADMIN	ESS Vision 2020 Implementation Project for Administrative Data
EU	European Union
EU-SILC	European Union Survey on Income and Living Conditions
FTP	File Transfer Protocol
GREG	Generalised Regression Estimator
GSBPM	Generic Statistical Business Process Model
GSDEM	Generic Statistical Data Editing Models
HBS	Household Budget Survey
HLFS	Household Labour Force Survey
IT	Information Technology
Istat	Italian Statistical office
KAU	Kind of Activity Unit
KOMUSO	ESSnet project on the quality of multisource statistics
LBD	Longitudinal Business Database
LEED	Linked Employer-Employee Data
LFE	Linkable File Environment
MAE	Mean Absolute Error
MEMOBUST	Methodology of Modern Business Statistics
NACE	Statistical Classification of Economic Activities in the European Community
NSI	National Statistical Institute
RAS	Iterative scaling method for balancing the entries of a non-negative matrix
SAE	Small Area Estimation
SBS	Structural Business Statistics
SHIW	Survey on Household Income and Wealth
SLA	Student Loans and Allowances
UN	United Nations
UNECE	United Nations Economic Commission for Europe
URS	Statistical business register (Austria)
URV	Administrative business register (Austria)
WP	Work Package

1 Introduction

1.1 Motivation for the guidelines

These guidelines on the use of estimation methods for the integration of administrative sources should help to:

- Promote the use of administrative data in the statistical production process;
- Promote the use of advanced methods in statistical data integration;
- Promote knowledge transfer about the use of administrative data in the ESS;
- Harmonise the methods used for administrative data in the ESS.

The guidelines are aimed at all kinds of statistical production of multisource statistics in different statistical domains, in particular demographic and social statistics and economic statistics.

The guidelines present a process-oriented description of methods, which facilitates the use of the methods in the statistical production process.

A glossary of the terminology used in the production is available, helping to improve user understanding of the use of administrative data. Furthermore, understanding the concepts for the evaluation of quality in multisource statistics would help in application of the guidelines.

1.2 Scope of the guidelines

The guidelines cover a number of important methods in the production of statistics from data sources which are a combination of administrative data and statistical data (Multisource statistics). The main emphasis is on the application of methods for the sub-processes “Data integration”, “Editing & and imputation”, and on the alignment (harmonisation) of statistical units and statistical variables when administrative data are used. Other well known-known statistical methods are interpreted in the context of administrative data. Furthermore processing workflows for the production of statistics using administrative data, the production of statistical registers, and the maintenance of registers are considered.

The application of the methods is formulated as usage scenarios for processing which can be executed by different methodological options. The application of the options depends on the application domain, the available data sources, the legal and organisational background of the NSIs, and the available resources. If a decision tree for the application of the options is known the decision tree is presented in the Guidelines. In a number of usage scenarios it is rather difficult to rank the options by universally accepted criteria. In such cases the guidelines propose, instead of the ranking, a number of indicators for the evaluation of the application.

The evaluation criteria can probably contribute to an improvement of the processes by discussing improvements with the data owners. An example of such improvements is a more efficient system for identification of units in administrative data.

1.3 Costs and risks

In the long run, using administrative data can reduce production costs. Its use also reduces the response burden for citizens.

The necessary prerequisite is to set up a structure which supports the use of administrative data in production. This structure includes legal and administrative issues like cooperation agreements with the administrative data holders, but also the necessary IT infrastructure for using the methods, and training of personnel for the new production processes.

Besides the advantages of using administrative data, there is a certain risk in the dependency of a NSI on the data holders, but this risk seems to be rather negligible as long as the infrastructure of the public administration is stable and policy makers rely on statistical information for decision-making.

1.4 Background to guidelines and basic definitions

The leading principle of the guidelines is the Generic Statistical Business Process Model (GSBPM). It is shown how the traditional survey oriented formulation of the GSBPM can be adapted to the use of administrative data. The first four phases of the GSBPM are only sketched in an introductory chapter. The main emphasis is on the estimation methods which are essential for the production of multisource statistics in the GSBPM phases 5 and 6.

Besides the GSBPM, a number of other sources are important. For the over-arching process of quality management in statistical production using administrative data, the results of the KOMUSO project about quality in multisource statistics is an important resource. The Guidelines will not repeat the details but a careful study of the project documentation is a necessary prerequisite for production with administrative data.

For most of the methods used one can find a more detailed description in a number of ESSnets. In particular the documentations of the following ESSnets are important sources for the presentation:

- ESS Vision 2020 Admin project (KOMUSO, WP 2),
- ESSnet on data integration,
- ESSnet on the use of administrative data and accounts data in business statistics,
- ESSnet on modern business statistics (MEMOBUST).

Besides this project, a valuable general reference is the book: A. & B. Wallgren (2014): Register-based Statistics – Administrative Data for Statistical Purposes.

The basic definitions used in the production of multisource statistics can be found in the Glossary of Work Package 7 of the ESS.VIP Admin project.

2 Use of administrative data in statistical production

This chapter explains a number of topics which are important for supporting the use of administrative data in statistical production and for giving the readers an understanding of the similarities and differences between statistical data and administrative data specificities.

2.1 Statistical data and administrative data

Statistical data are defined as data informing about statistical units, which are the objects of statistical surveys and the bearers of statistical characteristics. Usually statistical units are defined on the basis of three criteria: (i) legal, accounting or organisational criteria; (ii) geographical criteria; (iii) activity criteria. Important examples in economic statistics are the enterprise, the local unit, the enterprise group, employees, geographical areas, or events. In social and demographic statistics important units are the natural person, household, family, or living quarter.

Administrative data are defined in a similar way as a set of units and data derived from an administrative source. They are collected for the purpose of carrying out various non-statistical programs by institutions belonging to the government sector or by private organisations. Typical examples are registration of particular events like births and deaths, data for administration of social benefits, data about the taxation of individuals or businesses, or payments by individuals to business (for example for electricity or water consumption).

From the definitions, it is obvious that both types of data often capture similar economic or social phenomena and that the use of administrative data in statistics has a number of benefits. The most important benefits are: (i) a lower response burden; (ii) lower production costs; (iii) more information about economy and society; (iv) Increased availability of data; (v) more detailed information (for example about geographical areas or economic activity). Realisation of these benefits must resolve the differences between statistical data and administrative data. The most important differences are the following ones:

- Usually data collection of statistical data is organised in such a way that all necessary information about the statistical units is collected in a census or a sampling survey. Hence, the information is collected in one homogeneous environment. In administrative data the required information is often distributed over different administrative sources;
- The target population of an administrative source can be different from the target population for the statistical output. Hence, the units are collected over different administrative sources;
- The definitions of the administrative units are often slightly different from the definitions of the statistical units. Hence, alignment (harmonisation) of the units is often necessary;
- Administrative data often use concepts for the measurements of the characteristics which are different from the concepts of statistical variables and hence the alignment of variables is often needed;
- The definitions, the concepts and the target population of the administrative sources can

change over time. Hence, one of the most important tasks for the NSI is to control these changes. The best way is to define an official agreement with the holder (owner) to gather information about these changes;

- Administrative data have a different error structure in the measurements. The random sampling error is usually not meaningful for administrative data but the measurement error encompasses besides a random error a number of different bias components. The most important ones are: (i) Non-response error caused by missing information in the administrative data; (ii) Processing error arising from faulty implementation of planned methods; (iii) Coverage errors due to over-coverage or under-coverage of the population represented by the administrative source; (iv) Model error due to the use of other measurement methods.

The intention of these guidelines is to introduce a number of possible usage scenarios for the use of administrative data in statistical production and to highlight some methods and procedures which are useful in the use of administrative data and statistical data.

2.2 Usage of administrative data in statistical production

The Generic Statistical Business Process Model (GSBPM) defines a unified framework for statistical production. The formulation of the GSBPM emphasises traditional statistical production based on sampling data but the formulation of the different processing phases and the sub-processes in the phases also apply for other data. Consequently, the Guidelines rely on this processing model and show how the different phases and steps can be adapted to the use of administrative data in statistical production. As mentioned above, one important feature in the use of administrative data is the fact that usually several different sources are used rather than just one. This justifies the term multisource statistics for this type of statistical production.

Generally speaking, phases 1 and 2 are of a more conceptual nature and set the scene inside a NSI for the statistical production. In the case of administrative data, in phase 1: “Specify Needs” a number of sub-processes are more or less obvious because the administrative data should be used for the production of already existing products, e.g. a register based census. The most important are sub-processes 1.4: “Check data availability” and 1.6: “Prepare business case”. The former basically means to make an inventory of the existing administrative data which can be used, whereas the latter concerns the adaptation within the organisation inside a statistical office.

In GSBPM phase 2: “Design”, the most important issue is sub-process 2.3: “Design collection”. The necessary activities encompass a cooperation agreement with data owners (including in the ideal case technical and conceptual support for the use of administrative data), the organisation of data provision for administrative sources (including also responsibilities in the participating organisations), the development of a data repository for the administrative sources (data warehouse for administrative data inside the NSI), as well as all issues in connection with new challenges in data privacy and data security. The sub-process 2.2 “Design variables descriptions” has to be seen in connection with the development of a metadata repository for the administrative data. The sub-process 2.6: “Design production system and workflow” has to include the development of knowledge about peculiarities in the methods, which occur in connection with the production with administrative data. The subsequent chapters of these Guidelines will point to some important aspects.

GSBPM phase 3: “Build” is the transition from the more conceptual design of a new production to building a production solution. Sub-process 3.1: “Build collection instrument” refers to the specification of the physical data exchange between a NSI and the administrative data holders and requires an efficient and secure solution for data exchange, for example using FTP or EDI. Of utmost importance in the phase are the sub-processes 3.5: “Test production system” and 3.6: “Test statistical business process”.

If the organisational and administrative prerequisites are solved in a NSI, the data collections in the GSBPM phase 4: “Collect” is in some sense rather simple for administrative data. A challenge is the definition of a data model which organises all the information of the different sources.

The subsequent chapters will consider mainly guidelines for the GSBPM phases 5 and 6.

A number of activities in the ESS have already discussed in detail the necessary activities in connection with phases 1 and 2. In particular, one should refer to the reports of the ESSnet on the use of administrative data and accounts data in business statistics which contain a lot of useful information.

Furthermore, it should be mentioned that the over-arching process of quality management for the production based on administrative data has been analysed extensively in the KOMUSO project about quality in multisource statistics. The Guidelines will not repeat the details but a careful study of the project documentation is a necessary prerequisite for production with administrative data.

2.3 A typology of methods for administrative data

Types of methods for administrative data

Generally speaking three different types of methods can be distinguished in the use of administrative data.

1. **Type A methods**: Methods for sub-processes of the GSBPM for statistics based on administrative data;
2. Integrate data (GSBPM 5.1);
3. Edit & Impute (GSBPM 5.4);
4. Alignment of statistical units and measurements (GSBPM 5.2, 5.6);
5. Calculation of aggregates for multisource statistics (GSBPM 5.7);
6. **Type B methods**: Sub-processes of the GSBPM using administrative sources as auxiliary variables;
7. Using administrative data for weighting (GSBPM 5.6);
8. Using administrative data for calculation of aggregates (GSBPM 5.7);
9. **Type C methods**: Methods which define a production workflow with administrative data based on various GSBPM sub-processes;
10. Workflows for producing statistics from administrative data;
11. Workflow for producing statistics from administrative data and survey data;
12. Workflow for creation of statistical registers;
13. Workflows for updating and maintaining statistical registers.

Structure of the data used by the methods

The application of these methods depends on conditions on the data. The ESS Vision 2020 ADMIN project defines a number of data configurations for two datasets which occur frequently in connection with the production of multisource statistics. In the case of more than two datasets the configurations can be applied sequentially in the integration process. These configurations use a number of structural properties of the data based on the following attributes:

1. Structural conditions on the data:
   - Level of aggregation: micro-data, macro-data;
   - Temporal structure: time series (longitudinal data).
2. Combination of structural conditions:
   - Both datasets micro-data;
   - Both datasets macro-data;
   - Combination between micro-data and macro-data.
3. Representation of the envisaged population:
   - Completeness of the population: complete or incomplete;
   - Overlap of the population: disjoint or overlap of the population in the datasets.
4. Representation of the variables in the datasets:
   - Unique representation (variables only in one dataset) or multiple representation;
   - The values of the variables cover the total population or have missing information for some subpopulations.
   - Error in the variables or error-free variables;
   - The descriptions of the methods in the guidelines follow the above defined typology. For each method a short description of the problem is given, afterwards the different methods are described and a workflow for the application is outlined. For the workflow, reference to the corresponding GSBPM sub-processes is mentioned. After this general description, the usage scenarios are described by specifying the structure of the data used and the methodological options for the scenario are presented. Only those structural elements which are important are listed. The options are evaluated according to the preferences for use. Finally, an example is given and references for more detailed information.

General references for type A and type B methods are WP 2 of the ESS.VIP ADMIN project and the Memobust Handbook. For Type C methods the book of A. & B. Wallgren (2014) is a useful resource.

References

- Di Zio, M., Zhang, L.-C. and De Waal, T. (2017) *Statistical Methods for Combining Multiple Sources of Administrative and Survey Data*, The Survey Statistician (July 2017), 17-26
- Kish, L. (1965). *Survey Sampling*. Wiley.
- The Memobust Handbook of Modern Business Statistics (2014)[1]
- Myrskylä, P. (2004). *Use of Register and Administrative Data Sources for Statistical Purposes*. Statistics Finland.
- Reid, G. and F. Zabala and A. Holmberg (2017). *Extending TSE to Administrative Data:A Quality Framework and Case Studies from Stats NZ*. Journal of Official Statistics 33(2), 477–511
- Särndal, C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley and Sons, New York.
- Särndal, C.-E., Swensson, B. and J. Wretman (1992). *Model Assisted Survey Sampling*, Springer, New York.
- Wallgren, A. & Wallgren, B. (2014). *Register-based statistics: Statistical methods for*

[1] https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

administrative data. Wiley.

- Zhang, L.-C. (2012). *Topics of statistical theory for register-based statistics and data integration.* Statistica Neerlandica 66(1), 41–63.

3 Alignment of units and measurements

This chapter considers methods for which administrative data play not only a supporting role but are an essential part of the processes. The definition and the harmonisation of statistical units are practically impossible without the use of administrative data; for operational use, statistical units must reflect entities in economic and socio-demographic reality. The operationalisation of such units often depends on measurements defined by administrative and legal regulations. This dependence on existing administrative units often requires application of conceptual data modelling methods. With respect to the alignment of measurements, two basic scenarios are alignment by harmonisation of the value domains (classifications) used and alignment of measurements with errors.

3.1 Using administrative data for alignment of statistical units

Description of the problem

A statistical unit is an entity about which information is sought and for which statistics are ultimately compiled. It is the unit at the basis of statistical aggregates and to which tabulated data refer. In general, the definition of the unit encompasses a verbal definition of a number of important variables which are used for more detailed and formal characterisation of the units. Usually the value domain of these variables is standardised by international classifications (e.g. NACE).

For many important units, standardised international definitions are given (for example in connection with businesses the UN definitions for Units (2007)[2] or the UNECE recommendations for the Census 2020). However, sometimes the definition of new statistical units aims at capturing new economic and societal phenomena. An example can be ICT, where the description of the rapid development requires new units, like companies specialised in providing emerging new data services.

Administrative data often contain a lot of information about the statistical units. However, this information is given for administrative units which are not exactly the intended statistical units. The task of alignment of units is mapping the information about the administrative units onto the envisaged statistical units.

Methods and workflow for alignment of statistical units

Methods for alignment of statistical units

The methods combine statistical techniques for data integration with methods for data modelling from computer science. Strictly speaking, the definition of statistical units is more a conceptual task and

[2] See also Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community

can be done without using any data. However, for practical applications it is necessary that the population defined by the statistical units is provided in statistical data, in most cases a statistical register.

The methods for alignment of statistical units use the following modelling techniques:

- Conceptual data modelling from computer science;
- Data integration techniques which allow integration of data from different data models;
- Mathematical and logical data transformation techniques;
- Methods for documentation of database design.

General workflow for alignment of statistical units

This is a more detailed description of the GSBPM sub-processes 5.3: “Review & validate” and 5.5: “Derive new variables & units”.

The methodology for the alignment of the statistical units is based on the methods of data modelling from computer science and encompasses a number of steps. At first, a data model for the available administrative data is developed, allowing for identification of the legal units which are of interest for the statistical unit. Next, a pivotal administrative unit is identified by finding the unit which is conceptually as close as possible to the intended concept of the statistical unit. In a next step, the attributes of this pivotal legal unit and the relation to other legal units is investigated. From this structure, derivation rules for augmenting the variables of the administrative units are derived. Based on this analysis the variables for the administrative units can be transformed into variables for the statistical unit. An example how one can proceed in the case of the definition of enterprises is given in the Memobust handbook under the Theme: Derivation of Statistical Units.

This general case is often simplified if there exists a simple relation between the statistical unit and the administrative unit and the mapping is obtained by standard operations. One standard operation is specialisation of existing administrative units by additional attributes. The population of statistical units can be understood as the definition of a subpopulation of the administrative population. Other possibilities are the definition of a super-population by generalisation. From the methodological point of view, the definition starts with the summarisation of data into one dataset based on the common unique identifier. Harmonisation of the variables according to the definition of the statistical units is necessary (see 5.2).

Remarks on the alignment workflow

Obviously, the realisation of this workflow requires experience in the different techniques and a lot of domain knowledge.

Crucial points in the alignment of units are:

- Legal and administrative units are usually dependent on the national legislation, but units used across the ESS must be defined in the same way by all Member States;
- The alignment of units depends on some temporal stability in the production of administrative data. Any changes in the concepts used for the administrative data, as well as in the data maintenance, must be documented;
- The birth and death of the administrative units must be accurately documented.

Usage Scenario: Alignment of statistical units

Input

An analytical concept of the envisaged statistical unit; a number of administrative micro-data sources for administrative units which are related to the envisaged statistical unit. For the administrative data sources, a conceptual data model is known. Furthermore, the administrative sources cover the entire population of interest.

Output

A statistical dataset for the statistical units of interest.

Methodological options

Option 1: Augmentation of the data model for the administrative data with the concepts for the envisaged statistical unit and define the necessary data transformations for alignment.

Option 2: Use the existing data model for the administrative data and apply standard operations like specialisation or aggregation.

Option 3: Alignment of the units in an ad-hoc way.

Evaluation of the options

- If possible use Option 2;
- Option 1 should be used only in exceptional cases;
- Option 3 is not recommended.

No matter which option is used, documenting the transformation in the metadata of the new dataset is of utmost importance.

Examples

A rather complex application of the alignment of units can be found in business statistics, where different administrative units have to be aligned. For example, an enterprise may occur in administrative data in different guises: as a unit according to commercial law, as a unit according to the income tax, or as a unit according to the value added tax. In social statistics, persons also occur in different guises, for example as a person in the central population register or as an insurance case in the social security system.

Sometimes the alignment of statistical units is rather simple. Typical examples are the specialisation of already existing units by using a number of variables (attributes) for the units. A typical example is the definition of the statistical unit “person” in the labour force. The reverse operation to specialisation is aggregation, which defines units as the aggregation of different units. An example of such a hierarchical relation in business statistics are the following units: a business group is the aggregation of a number of businesses, the business itself may be the aggregation of a number of Kind of Activity Units (KAU) ,and the KAU itself may be the aggregation of Local Kind of Activity Units.

References

- Eurostat (2010): *Business registers, recommendations manual*, Luxembourg.
- United Nations (2007): *Statistical Units*, New York.
- van Delden, A., Lorenc, B., Struijs, P. and Zhang L.-C. Letter to the editor. On Statistical Unit Errors in Business Statistics. Journal of Official Statistics, Vol. 34, No. 2, 2018, pp. 573–580

3.2 Alignment of measurements for administrative variables

Description of the problem

The GSBPM sub-processes 5.2, 5.3 and 5.5 encompass the necessary steps for modifications of variables in survey data. In the case of administrative data, there are two main reasons for modifications of the variables in the administrative dataset. The first one is the case when administrative variables use a different concept and the derivation of a new variable is necessary. This corresponds to the sub-process 5.5: “Derive new variables & units”. The second reason is that the administrative variable uses the correct concept but the values of measurement are different. This corresponds to the sub-process 5.2: “Classify & code”.

Similar to the case of statistical variables, one has to be aware that administrative information may be flawed by processing errors in the administrative data. Two typical reasons for errors in administrative data are the fact that the administrative variable is of minor importance for the owner of the administrative data, and the so-called administrative delay caused by the fact that events are recorded in administrative data with some delay. Correction of such errors in administrative variables corresponds to activities in the sub-process 5.3: “Review & validate”.

The term “alignment of measurements” can be understood as an umbrella term for these three cases.

Methods and workflow for alignment of measurements

Methods for the alignment of measurements

Corresponding to the similarities of the tasks in the GSBPM sub-processes, the methods which can be applied for the alignment correspond to the methods used for survey data. The most important are:

- *Supervised methods*

  These are all methods that model the relationship between the independent variables (covariates) and the target variable. We assume that the target variable we model is the target measure we would like to be aligned. They can be used for the derivation of new variables, as well as for classification and coding. Using the knowledge about the relationship between the administrative data and the intended target variable, a transformation which allows the estimation of the values for the intended variable is defined. Typical examples of such functions are the recoding functions in case of categorical variables or regression in the case of numerical variables. Evaluation of the transformation can be done if statistical data are available. In the case of transformation by regression, the existence of survey data for the variable is necessary for estimation of the regression coefficients. The quality of the model is evaluated by the usual diagnostic instruments for the regression.

- <u>*Unsupervised methods*</u>

  These are all methods that treat the measurements without any hierarchy, i.e. that assume that all the measurements are in principle close but not necessarily the same as the true target value.

  Latent class models are used for estimation of categorical variables. In this case, the categories of the variable are interpreted as class indicators and the observed classes are interpreted as the outcome of a latent class model. The class itself is the latent target variable indicating the true value of the class membership for the data. As in the case of standard models, a statistical dataset containing the statistical variable for at least some statistical units is necessary for specification of the model. This method can be used also for review and validation if the statistical data are not error-free.

  Structural equation models are used for estimation of numerical variables. This approach is similar to latent class modelling, but assumes that the target unknown variable is a numeric variable and the result of a structural equation model. Usually, a model is used which assumes that each observed variable in the administrative data can be modelled as measuring one latent target variable. For identification of the parameters a random free subsample of the target variable is necessary. This method can be used also for review and validation if the statistical data are not error-free.

General workflow for alignment of measurements

The workflow is defined by the GSBPM sub-processes 5.2: “Classify & code” and 5.3: “Review & validate”, taking into account peculiarities of multi-source statistics.

The alignment of measurements should use the general ideas for the workflow in statistical modelling. The following steps should be performed:

- Record matching for integration of the administrative data and the statistical data used for model specification and validation;
- Evaluation of the quality of both sources, in particular the consideration of time periods, may be important for the statistical data used for validation;
- Considerations about pseudonymization in the case of confidential data;
- Application of the appropriate transformations;
- Evaluation of the results using the diagnostic tools for the method (for example residual analysis);
- Application of the transformation for all units in the dataset.

Remarks on the workflow for alignment of measurement

The workflow makes the hidden assumptions that application of the transformations is correct from a statistical methodological point of view (for example normal error terms for the regression). In all cases, a good documentation of the processes is necessary.

Usage Scenario: Alignment of statistical variables

Input

Two or more data sets linked at micro-level, containing the target variable for alignment and the explanatory variables used in the transformation model. The target variable can also be observed only for a subset of data.

Output

A rule for alignment of the measurement of the target variable, which can be used in further processing; application of this rule for the administrative dataset leads to a dataset with the intended statistical variable.

Methodological options

- Option 1: Supervised methods.

- Option 2: Unsupervised methods.

Evaluation of the options

- When it can be assumed that in one data set (either administrative or sample survey) the observed target variable reports the true values, option 1 (**supervised models)** should be used. In such circumstances, a model considering this variable as response variable and the other target variables to be aligned as covariates, should be estimated. The estimated model, applied to the observed values of the covariates, aligns those variables to the target true one.

- When we cannot assume that one variable is the true one, option 2 (**unsupervised approach)** should be used. In practice, the true target variable is considered as a latent variable, and the observed target variables are considered as imperfect measurements of the true target variable. The objective is to predict the true value given the observed values of the target variables (proxy) observed in the data sets.

Examples

In the ESS Vision 2020 ADMIN project, Work Package 2 has applied structural equation modelling for estimating the turnover from tax data (WP2, Task 2.1.1, Template T-4.1) and latent class modelling for estimating the classification error for the variable “home ownership” (WP2, Task 2.1.1, Template T-4.2).

References

- ESSnet on data integration, WP1.
- Oberski, D. (2017), *Estimating Error Rates in an Administrative Register and Survey Questions Using a Latent Class Model.* In: Biemer, De Leeuw, Eckman, Edwards, Kreuter, Lyberg, Tucker and West (eds.), Total Survey Error in Practice: Improving Quality in the Era of Big Data, New York: John Wiley & Sons.
- Oberski, D. L., Kirchner, A., Eckman, S., & Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*, 112(520), 1477-1489
- ESS.VIP.ADMIN, WP2 Summary sheet on estimation methods 6b.

4 Use of administrative data in editing and imputation

Administrative data can be used in editing and imputation in two different usage scenarios:

- Administrative data as an additional micro-data source together with statistical data as input for the editing and imputation process;
- As additional source of auxiliary information for editing and imputation of statistical data.

Sections 4.1 and 4.2 consider these applications. Each section is organised in the same way: after a description of the problem, the methods and the workflow for editing and imputation are presented and peculiarities in connection with the use of administrative data are outlined. Afterwards the usage scenarios and the methodological options for the usage scenarios are described and an evaluation of the alternatives is given. At the end of the section references for a more detailed description are mentioned.

4.1 Using administrative data for editing of statistical micro-data

Description of the problem

Unavoidably, data collected by the Statistical Institutes, as well as data from external sources used in the statistical production process, are not error-free. In order to ensure that the final statistical product is of good quality, it is crucial to detect and treat such errors at an early stage of the statistical production process. In survey data the main source of errors are the random sampling error and the measurement errors, resulting from all the activities and the different actors in the statistical production process. In the case of administrative data there is usually no sampling error but random measurement error and different kinds of bias, for example bias due to the differences in the definitions between the concepts of the administrative data and the statistical data.

Administrative data can play an important role in the editing process: (i) it can add information to the statistical data, or (ii) it can be used as a reference dataset. Administrative data can support the editing task by helping in error localization, in confirming the values of variables that do not satisfy the set of edit-rules and by contributing to the improvement of the detection of erroneous variables. Despite all evident advantages of using administrative data in this process, it should be emphasized that administrative data suffer from the same type of errors as statistical data do: missing items or missing records, measurement and processing errors. Additionally, the lack of quality control over the data can be a limitation when using these data and may imply extensive additional data editing activities as well as the need of checking the consistency and/or completeness of the data before its use.

The workflow is a detailed description of the GSBPM sub-processes 5.3: “Review & validate” and 5.4: “Edit & impute”.

Methods for editing micro-data

Methods for editing

For a given micro-dataset several procedures for data editing can be used. The most important ones are:

- Manual (interactive) editing: This is the traditional approach to check and correct micro-data. It requires that the identification of the units that need a closer inspection should be done by an expert in the subject and population, supported by software for interactive editing. This justifies the name interactive editing for this method.
- Deductive editing: Deductive editing is a procedure editing systematic errors, i.e. errors for which the error mechanism is known. Typical examples are sign errors, typing errors, errors in the unit of measurement, or violations of obvious rules for the values. In the latter case the hidden assumption is that some of the variables defining the rule are error-free.
- Automatic editing: Automatic editing is a method used for the detection of random errors after the correction of all systematic errors by deductive editing. It is applied mainly for numeric variables by formulation of edit rules as mathematical equations.
- Selective editing: This method detects potentially influential errors in numeric variables. Influential errors are defined as errors in values of variables that have a significant impact on the statistics produced.
- Macro-editing: This method differs from the Selective editing method only in the form that units are selected. Selective editing methods use data of a single unit and related auxiliary information to determine the possibility of the presence of an influential error. Macro-editing selects units by considering all or a large part of the data. Therefore this method can only be considered if a substantial part of the data is available.
- Outliers and extreme values detection: The objective of these methods is the identification of the outliers in numeric variables. For outliers one has to decide whether the value is an error or an influential correct value. Other methods for detecting outliers are at the beginning by selective editing or by macro-editing at the end of the editing phase.
- Methods for reconciling conflicting microdata: When using information from different sources, the composite records may consist of several combinations of sources. The combination may give rise 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 reconciling conflicting microdata is to solve the consistency problems by making small adjustments to some of the variables involved. The values in the record with inconsistent microdata are changed, as little as possible, such that the modified record is consistent in the sense that it satisfies all edit rules. Methods: prorating, minimum change adjustment, generalised ratio adjustment. (Further details can be found in Work Package 2 of the ESS Vision 2020 Administrative data sources project)[3].

[3] https://ec.europa.eu/eurostat/cros/content/task-211-estimation-methods-integration-administrative-sources_en

Usage Scenarios:

Scenario 1: Editing datasets for using variables from both datasets for statistical production

Input

In the case of multisource statistics data editing must be done for two or more micro-datasets. The data can be administrative data or statistical data. The editing can be used only for those units for which variables defining the possible edit rules are available (not necessarily the complete population but overlap in the populations allowing the formulation of edit rules). Integration of the datasets by using methods for record matching is possible (unique keys or matching variables). The variables of interest can have multiple representations in both datasets.

Output

A micro-dataset, containing information from both datasets, which is error-free but probably contains missing values that will be treated afterwards in the imputation step.

Methodological options

Design of a data editing process

Different workflows may be designed for different types of statistical production processes in terms of type of investigated units (enterprises, households), variables (continuous, categorical) and sources (direct surveys, integrated sources). In UNECE's Generic Statistical Data Editing Models (GSDEMs) ver. 1.0 (2015), developed on behalf of the international statistical community, generic workflows for structural business statistics, short-term business statistics, business census, household statistics and statistics through data integration are described. A workflow for administrative data may be obtained by combining the previous models.

Three options can be envisaged:

- Option 1: Each source is edited separately; afterwards, the sources are integrated. The integrated sources are edited once more in order to assess the consistency among the variables' values, obtained from the different sources.
- Option 2: Integrate the sources keeping only the variables that are of interest; afterwards, apply editing for the integrated dataset.
- Option 3: Apply a "light" editing for each single source, which considers at least all variables used for matching of the sources. Afterwards, integrate the sources and edit the integrated dataset.

Evaluation of the options

Option 1 is preferable but time/resource consuming.

Option 2 is less time-consuming but has the disadvantage that due to the possible loss of some variables not all edits are detected.

Option 3 is a compromise taking into account a trade-off between the quality of the resulting data and the resources available.

Descriptive information about the editing process should be included.

Useful indicators for the editing process are:

- Information for all variables on the share of edits;

  In the case of edit rules using variables from both datasets, the following indicators should be used:

  - Share of edits for all units for which deterministic matching is possible;
  - Share of edits for all units with probabilistic matching.

One can define thresholds for these shares and in this way define quality criteria for the datasets. For example: 1 = “good, eligible for data production”, 2 = “eligible for data production”, 3 = “not eligible for data production”.

Evaluation of methods

Methods are chosen according to the errors they are intended to treat, the nature of the variables, etc. (see Edimbus, Memobust and UNECE 2015). The short description of methods previously given is clear regarding their application. However, editing of administrative data presents some peculiarities which are worthwhile mentioning, in order to choose methods.

- It is important to take into account the quality of the administrative data. Most of the methods used in data editing allow taking into account the different quality of the variables involved in the editing step. For instance, in the Fellegi-Holt algorithm, a weight proportional to the reliability of the variable can be introduced in the algorithm, so that the more reliable variable is less frequently changed.
- Automatic editing. The Fellegi-Holt paradigm, based on the minimum cardinality of variables to be changed in order to make an erroneous unit consistent with respect to the edit rules, cannot always be adopted with administrative data. This paradigm implicitly assumes that the error is random. When using administrative data, the error should be given by problem of alignment of measurement, data collection, small difference in the definition, and so forth. In such a case, there is a systematic difference. In those cases, methods for reconciling conflicting micro-data should be used.
- Selective editing presumes the possibility of recontacting the units. This is generally not feasible when using administrative data. When using selective editing in this context, we need to take into account this problem.
- See other suggestions in Memobust 2014.

Scenario 2: Editing of variables if administrative data are used for auxiliary information

Input

A micro-dataset of primary interest and an administrative dataset containing auxiliary information about the variables in the micro-dataset of primary interest. The administrative dataset may be micro-data or macro-data.

Output

An error-free version of the micro-dataset of primary interest, which is consistent with the information in the administrative dataset.

Methodological options

Design

The situation is similar to the previous discussion. However, the role of the administrative data are not as prominent. In such a case, the idea of first checking the data sources separately and then checking the integrated data is generally considered not suitable and much too expensive. The workflow is thus generally reduced to the ones concerning single surveys, i.e. the previous option 2: Integrate the sources keeping only the variables which are of interest; afterwards, apply editing for the integrated dataset.

Methods

See previous subsection on methods

Examples

The Memobust handbook provides a number of examples for the different edit methods.

In the case of register based census the principle of redundancy is often used. This means that a variable of interest occurs in more than one dataset. Based on an evaluation of the quality of the variable in the different datasets one can define edit rules for such variables. In the case of more than two occurrences a simple rule can be editing the variable according to the most frequent occurrence of the values.

References

- De Waal, T. Pannekoek, J, and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- 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
- The Memobust handbook (2014), Module Statistical Data Editing - Editing administrative data.
- Särndal, C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley and Sons, New York
- UNECE (2015), on behalf of the international statistical community, *Generic Statistical Data Editing Models (GSDEM)* Version 1.0[4]
- Wallgren, A. & Wallgren, B. (2014). *Register-based statistics: Statistical methods for administrative data*. Wiley.
- Zhang, L-C. (2015). *On modelling register coverage errors*. Journal of Official Statistics, 31(3), 381-396.
- Zhang, L-C. (2011). *A unit-error theory for register-based household statistics*. Journal of Official Statistics, 27(3), 415-432.

[4] https://statswiki.unece.org/download/attachments/117771706/Generic%20Statistical%20Data%20Editing%20Models%20v1_0.pdf?version=2&modificationDate=1450253100949&api=v2

4.2 Using administrative data for imputation of statistical micro-data

Description of the problem

Imputation is the process of replacing the missing values with estimated values from the available data. In surveys, missing data occur when respondents are not willing or able to provide answers to one or more questions in the survey. In administrative data, missing values can occur due to incomplete data capture in the administrative source. Another source of missing values is the editing process when potentially erroneous/conflicting values are detected. It is often common practice to replace them by missing values. In the imputation phase, these missing values will be replaced by plausible ones in order to guarantee data quality necessary for analysis and dissemination.

The consequences of the presence of missing data are already well understood: (i) it reduces statistical power, (ii) it can cause bias in the estimation of parameters, (iii) it can reduce the representativeness of the samples, and (iv) it may complicate the analysis of the study. Each of these distortions may threaten the validity of the results and can lead to invalid conclusions. Therefore, the use of imputation methods is crucial to guarantee the quality of the micro-data and/or of parameter estimates.

There are several procedures to handle missing data and the application depends on the type of dataset, its scope, the nature of the missing values, and on the objectives.

Regardless of the imputation method considered, at the end of the procedure the consistency of the imputed observations should be checked and the impact of imputations on estimates and variances should be evaluated.

The imputation methods presented in this section are only used for item non-response. Unit non-response will be treated in the weighting process and mass imputation (e.g. blocks of missing values) will be treated in univalent estimation in chapter 7.

The workflow is a more detailed formulation of the GSBPM sub-process 5.4: “Edit & impute”.

Methods and workflow for imputation

Methods for imputation in micro-data

The following method will be considered:

- Deductive imputation: This method identifies cases where it is possible, based on logical or mathematical relationships between the variables, to unambiguously derive the value of one or more missing variables from the values that were observed.
- Model-based imputation: Univariate methods find a predictive model for each variable that contains missing values, as it is the case of regression imputation, mean imputation or ratio imputation. The auxiliary variables used to construct the predictive model may be obtained from the current survey or from other sources (historical information or administrative data). Nevertheless, it should be stressed that these methods could lead to bias in the estimation of the relation between the variables. If that bias is not negligible then multivariate approaches should be used.
- Donor imputation: The objective of these methods (cold deck, random hot deck, sequential hot deck, nearest neighbour, etc.) is to replace the missing value of a given unit by copying the observed value of another unit, the donor. Typically, the donor is chosen in such a way that it resembles the imputed unit as much as possible on one or more background characteristics.

  - Some methods do not take into account the constraints that the statistical data must satisfy. Thus, there is no guarantee that the imputations made will satisfy the edit rules. A two-step procedure is generally adopted. First, the missing values are imputed without taking (all) constraints into account. In the second step, the imputed values are minimally adjusted to satisfy the edit rules, according to some criterion (see reconciling conflicting microdata in section 3.1).

Usage scenario: Imputation of multisource data defined by statistical data and administrative data

Input

A micro-dataset probably obtained after the integration of different micro-datasets for which editing was applied. The dataset should represent the entire population and the variables should be error-free.

Output

A micro-dataset in which missing values have been replaced by imputed values.

Methodological options

Design for imputation of micro-data

- When using integrated administrative data, it is important to evaluate the quality of the different sources. Different level of quality may suggest different methods, for instance robust modelling.
- When imputation is performed, the first step is to think about the missing data mechanism. All the usual methods are based on the Missing at Random (MAR) assumption, while administrative data may contain systematic missing units or values in some part of the population (strata of the population not observed). This systematic mechanism is even more evident when administrative data are integrated with other sources, for instance a sample survey. In these cases, the usual methods may introduce a bias in the estimates. Covariates explaining missing data should be used in the imputation methods in order to adjust for the bias.
- If applicable, deductive imputation has preference over all other imputation methods. Nevertheless, this method requires an error-free dataset. For this reason, it should be applied only after extensive editing has been performed.
- If multiple missing values must be imputed in a single unit, then donor methods are easier to use and might preserve, as much as possible, the correlations between the variables. Nevertheless, it is not parsimonious, since it essentially estimates – roughly speaking – all interactions between variables, hence it requires a high number of observations.
- Imputation with or without a residual. Imputation with a residual, for instance the regressed value plus a random noise, is generally preferable when variability of the variable should be preserved. On the other hand, the introduction of a random residual increases the variability of the estimator compared to the one without a residual. In some cases preservation of the variability is not necessary for the target parameters, for instance when the target parameter is a total. In such a cases, an imputation without residual can be preferable.
- Finally, in conjunction with the application of the imputation method or later, after the application of imputation methods, the results must be checked and a decision made on whether an adjustment of the imputed values is necessary. Methods used for reconciling conflicting microdata may be used, such as ‘prorating’, ‘minimum change adjustments’, etc.

- In any case, it is strongly recommended to document the imputation by descriptive information. Beside the criteria mentioned in the remarks on the workflow, one can use the following indicators:
    - Share of missing values in the variables;
    - Distribution of the missing values in different subgroups of the population;
    - Correlation between the imputed variables and the variables used for imputation.

Evaluation of methods

- In order to choose an imputation method or a procedure composed of a set of methods, beside considering the properties of methods that are generally known, it is important to assess their performance through empirical evaluations based on experimental data resembling as much as possible real data.

  This empirical evaluation can use the comparison with historical data for which the missing values are known.

  Another method is using the principle of training and test sets, well known in machine learning. This means that a subset of the existing data, without missing values, is taken as a test set. In this test set, missing values are created according to a specific mechanism, e.g., MCAR or MAR. For these random missing data the imputation procedure is applied and performance indicators are calculated by comparing the imputed values with the observed values.

  The performance indicators are micro and macro.

  Micro data evaluation. Indicators frequently used for the assessment of prediction at micro level are the correlation coefficient (*R*) between raw data and imputed data, the index of agreement (*d*) and the mean absolute error (*MAE*). All of these are based on the differences between the predicted values and their corresponding observed values. In the case of donor imputation, the frequency of how often a certain donor is used is another criterion.

  Macro evaluation compares the target parameter to be estimated, e.g., the total of the population, computed on raw data and on imputed data. In these simulated cases, it is important to make a Monte Carlo evaluation by repeating the procedures several times and averaging the results over the repetitions. In the case of missing values in the explanatory variables, extensions of the model-based methods can be used, for example sequential regression imputation.

- In general, it is not preferable to use weighted imputation. If possible, it is better to include a variable that forms the basis of the weights in the model.

The use of imputation classes is an interesting alternative if there are strata in the data and the variations in the strata are small compared to the variations between the strata. In this case, the use of different imputation models can be useful.

Examples

General examples for the different imputation methods can be found in the references.

In a register based census it is frequently necessary to use specific methods for the variables. Furthermore, the sequence of imputation must be defined carefully. Using the experience gained from a feasibility study, the Austrian register based census defined a sequence for the imputation of the variables. The first variable for imputation was the ´month of birth'. The imputation was done by choosing simply a random number. Next, for people not born in Austria the arrival in Austria was estimated by regression imputation. Afterwards the family status was estimated using hot deck imputation. A rather complicated deductive imputation method was designed for the imputation of the variable “type of household”.

References

- The Memobust handbook, Module Imputation.
- De Waal, T. Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.
- Jae Kwang Kim, Seho Park, Yilin Chen, Changbao Wu, *Combining Non-probability and Probability Survey Samples Through Mass Imputation*

5 Use of administrative data in weighting and estimation of micro-data

In weighting and estimation additional data are used as auxiliary variables for calculation of weights or for adjustment of estimates. This section considers applications of administrative data in the weighing and estimation process. The first one is the application in connection with classical survey data and the second one is the use of administrative data in the case of small area estimation.

5.1 Using administrative data for weighting and estimation

Description of the problem

Weighting and estimation is the process of obtaining, from the observed sample data, estimates for the parameters such as totals, means and ratios. Although it is assumed that errors and missing values have already been eliminated in the editing and imputation process, it is well known that imputations may lead to biased parameter estimates. Even if globally the impact of the imputations is small, it may happen that this is not true for certain sub-populations.

The method in this section mainly concerns the use of external information in order to produce unbiased or approximately unbiased estimators in probability-based sampling designs. In probability-based sampling designs, a design weight is associated to each sampled unit. Macro-level weighting will be described later. Using auxiliary information based on administrative data, they can be modified for improving the quality of the estimates. Usually the auxiliary information is taken from administrative data but this is not necessarily the case.

Methods and workflow for weighting and estimation

Methods for weighting and estimation

Calibration is the technique to adjust weights according to auxiliary information. One frequently used calibration method is GREG, which minimise Euclidian distance:

- Generalised regression estimator – GREG. This method is used for estimation when auxiliary information is available at unit or domain level. GREG is a model assisted estimator designed to improve the accuracy of the estimates by using the relationship between the target variable and the auxiliary variables. Since the resulting weights allow the calibration of the known totals, it is in fact a special case of a calibration estimator when the Euclidian distance is used.

  It can be used to reduce the variance of the estimates if a strong correlation between the target variable and the auxiliary variables exists. Therefore, its use is recommendable only if a linear relationship between target and covariate variables is present. However, when there is information about the auxiliary variables for each unit of the population, it is possible to consider non-linear variants of the GREG estimator in the case of a markedly non-linear relationship between auxiliary variables and the target variable (for example, logistic GREG).

Workflow for weighting and estimation

The most important steps for calculating new weights based on auxiliary information are defined by the GSBPM sub-processes 5.6: “Calculate weights” and 5.7: “Calculate aggregates”.

- Apply Editing & imputation for the data for achieving error-free and complete data;
- Selection of appropriate auxiliary variables which are correlated with the target variables;
- Calculation of the new weights by adaption of the design weights according to the model defined by the auxiliary variables;
- Check the effect of the calibration on variance (e.g. by checking the Kish factor)
- Using the new weights for tabulation;
- Documentation of the process.

Remarks on the weighting and estimation workflow

Depending on the aggregation level of the auxiliary information, which is not observed in the survey, it may be necessary to integrate the auxiliary data and the survey data.

Usage scenarios

Input

Survey data at the micro level based on a sample design; auxiliary variables either at the micro level for all units in the population or as macro-data.

Output

Sample data with modified weights according to the model based on the auxiliary information.

Methodological options

Option 1: Use the auxiliary information at the aggregated level and modify the Horvitz-Thompson estimator according to the regression equation defined by the auxiliary variables for the target variables.

Option 2: Use the auxiliary variables at the unit level for prediction of the values of the target variables for each unit in the population. The GREG estimator is calculated as the total of the predicted values plus the sum of the difference between predicted values and observed values for the sample.

Evaluation of the options

The application of the options depends on the available auxiliary information:

If the auxiliary variable is available at the unit level, Option 2 is preferable because it gives more information;

If the auxiliary variable is available only at the macro level, use Option 1.

Examples

The Memobust handbook gives a detailed example of the GREG for the Small and Medium-sized Enterprises (SME) sample survey in Italy. From the business register, the auxiliary variables “Number of employees in the previous year” and “Total number of enterprises in different categories” are used.

References

- The AdminData ESSnet: Use of Administrative and Accounts Data for Business Statistics, WP3 - Methods of Estimation for Variables
- The Memobust Handbook of Modern Business Statistics, Module Weighting and Estimation, Method Generalised Regression Estimator.
- Deville, Jean Claude and Carl Erik Särndal (1992): *Calibration estimators in survey sampling*. Journal of the American Statistical Association. 87(418), 376 – 382.
- Särndal, Swensson and Wretman (1992), *Model Assisted Survey Sampling,* New York, NY, US: Springer-Verlag Publishing.
- Särndal and Lundström (2005), *Estimation in Surveys with Nonresponse,* John Wiley & Sons. Chichester, England
- Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185-193
- Haziza, D., & Beaumont, J. F. (2017). Construction of weights in surveys: A review. *Statistical Science,* 32(2), 206-226.
- Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J., & Thompson, M. (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, 32(2), 227-248.
- Breidt, F. J., & Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190-205.

5.2 Using administrative data in small area estimation

Description of the problem

In official statistics surveys are usually planned to produce estimates at a higher level. If one is interested in estimates for smaller domains, the sample size is not large enough (sometimes without any observations) to produce reliable estimates. Using small area estimation (SAE) it is possible to obtain predictors for each small area for the target variables of interest. Small area estimation is a model-based approach which calculates the estimates as best linear unbiased predictors for the different domains.

Two methods are considered in this section. The first one is estimation by the EBLUP which defines a linear mixed model and the second one is small area estimation for time series data.

Methods and workflow for small area estimation

Methods for small area estimation

The following methods are considered:

- Area Level for small area estimation – EBLUP. These methods are a set of techniques allowing the estimation of parameters of interest for domains where the direct estimators cannot be considered reliable enough (present a very high variance). Small area methods increase the reliability of estimation by “borrowing strength” from a set of areas in a larger domain for which the direct estimator is reliable. This means that information from other areas is used and/or additional information from other sources (in particular, administrative data) is exploited.
- Small area estimation methods for time series data. These methods extend the basic area estimation model to handle time series and cross-sectional data. The improvement of the quality of the estimates is achieved by introducing linear relationship between target and auxiliary variables and explicitly modelling time dependent parameters. The methods can be applied when auxiliary information is available either at the unit level (Time series unit level models) or at area level (Time series area level models).

Workflow for use of small area estimation

The most important steps for calculating new weights based on auxiliary information are defined by the GSBPM sub-processes 5.6: “Calculate weights” and 5.7: “Calculate aggregates”:

- Apply Editing & imputation for the data for achieving error-free and complete data;
- Selection of appropriate auxiliary variables which can be used as covariates in the estimation process;
- Specification of the model;
- Calculation of the predictors in the different areas;
- Evaluate the results;
- Documentation of the process.

Remarks on the workflow

The methods can be applied if no sample data are available for some areas.

The covariates must be known only at the domain level.

Model specification is of utmost importance for obtaining reliable results.

Usage scenarios

Input

Macro-data for the target variables and the covariates (often administrative data); if time series estimates are required, these data must be defined for each time step.

Output

Prediction of the small area estimates for each domain; in the case of time series application, also for each time period.

Methodological options

Option 1: Use EBLUP small area estimation.

Option 2: Use small area estimation for time series.

Evaluation of the options

The decision about the method depends on the data structure.

If only cross-sectional data are available use the EBLUP;

If time series data are available use the small area method for time series.

For evaluation, inspection of the correlation is necessary because the methods require a strong relationship between the set of covariates and the target variable.

Furthermore mean square error and bias of the estimates should be checked using diagnostic tools.

Another assumption is the normality assumption. It may be necessary to transform data before applying the methods.

Examples

Using the EU-SILC survey data and population census data from Italy, small area estimation was used to estimate poverty at the regional and sub-regional level.

References

- *Memobust Handbook on Methodology of Modern Business Statistics* (2014), Module Weighting and Estimation, EBLUP Area Level for SAE.
- *Memobust Handbook on Methodology of Modern Business Statistics* (2014), Module Weighting and Estimation, SAE Time Series Data.
- Rao, J.N.K. (2003): *Small area estimation*. John Wiley & Sons, Hoboken, New Jersey.
- Zhang, Li-Chun and Ray Chambers (2004): *Small area estimates for cross-classifications*. Journal of the Royal Statistical Society (Series B). 66(2), 479 – 496.
- Rao, J.N.K. and Molina, I. (2015). *Small area estimation*, Second Edition. Wiley, Hoboken (NJ)
- B.B. Khare, Ashutosh and S.Khare (2018). *Comparative study of synthetic estimators with ratio synthetic estimator for domain mean in survey sampling using auxiliary character*, International Journal of Applied Mathematics and Statistics Vol.57, Issue No. 3.
- Gelman, A. and Carlin, J. B. and Stern, H. S. and Rubin, D. B. (2006). *Bayesian Data Analysis*, CRC Press Company
- Pfeffermann, Danny. *New Important Developments in Small Area Estimation*. Statist. Sci. 28 (2013), no. 1, 40--68. doi:10.1214/12-STS395
- Zhang, L. C., & Fosen, J. (2012). A modeling approach for uncertainty assessment of register-based small area statistics. *Journal of the Indian Society of Agricultural Statistics*, 66, 91-104.
- Zhang, L. C., & Giusti, C. (2016). Small area methods and administrative data integration. *Analysis of Poverty Data by Small Area Estimation*, 61-82.
- Tzavidis, N., Zhang, L. C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish:

a framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4), 927-979.

- SAMPLE project, (2010), http://www.sample-project.eu
- Pratesi, M. (2016). *Analysis of poverty data by small area estimation*., Wiley -

6 Data integration for statistical micro-data and administrative data

From a technical point of view data integration is the most important activity in the statistical production process for obtaining extensive information about statistical units. The data sources for integration can be survey data as well as administrative data. The integration is often hampered by the fact that identification of the units in the different data sources is not straightforward and that there is no guarantee that the information in the different sources refers to the same statistical units.

A number of methods have been developed for overcoming these problems. The two sections of this chapter consider two broad classes of integration scenarios. The first one is linking of the units based on object characteristics and the second one is statistical matching or data fusion, where a synthetic dataset is built using partial information about units in different sources.

6.1 Methods based on matching of object characteristics

Description of the problem

The goal of these methods is to identify the same real world entity, at micro level, that can be differently represented in data sources, even if unique identifiers are not available or are affected by errors. In statistics, object matching (known commonly as record linkage) is needed for several applications, including: enriching the information stored in different datasets; deduplication of datasets; improving the data quality of a source; measuring a population amount by capture-recapture method; checking the confidentiality of public-use micro-data.

At the end, the matching of the objects characteristics 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. In the 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.

Object matching is a challenging problem because of: errors, variations and missing data in the information used to link records; differences in data captured and maintained by different databases, e.g. age versus date of birth; data dynamics and database dynamics as data regularly and routinely change over time (e.g. address changes, name changes due to marriage and divorce).

There are several procedures to deal with these problems and the choice of the most suitable procedure will depend essentially on the quality of the common objects that are in the sources.

Regardless of the matching method considered, at the end of the procedure some manual checks are required to be sure to avoid mismatches and missing matches.

Methods for integration based on object characteristics

Methods for matching object characteristics

The following methods can be used for the matching problem

- Unweighted matching of object characteristics. This record linking method is intended for matching two datasets on the basis of object characteristics. It is applied in case no object identifiers (of good quality) are available from both datasets. First the potentially matching records in the two datasets are identified. This requires a suitable metric and a cut-off value so that records that are too different are not considered as candidate matches. In the next step, from these potential matches a subset is computed that maximises the number of matches, under suitable constraints
- Weighted matching of object characteristics. This method is applied to match two datasets with many common units, on common object characteristics. The method is able to value the strength of possible (candidate) matches by using matching weights. Weighted matching can be formulated as an optimisation problem, in which the optimal (weighted) sum of matches is calculated, under certain constraints, such that each record can appear in at most one of the matches. The goal of the method is to find solutions to such problems, exact ones or good approximations.
- A special case of weighed matching, probabilistic record linkage covers a set of methods whose aim is to identify a statistical unit even if it is differently represented in different data sources. The sources are composed mainly of the same units. These methods use probabilities to individuate the same units in different datasets; moreover they allow to measure the linkage error and so to evaluate the performance of the procedure.

General workflow for matching of object characteristics

The workflow mentions some of the most important topics in the execution of GSBPM sub-process 5.1: “Integrate data”.

- Descriptive analysis of the datasets (coverage);
- Identification of object characteristics for matching;
- Selection of matching method;
- Documentation.

Remarks on the general workflow

The object characteristic values used in the matching may contain missing values; this can have negative influence on the matching performance.

To assure the data quality, editing is necessary (refer to the editing chapter 3.1). After matching also imputation may be necessary (refer to the imputation chapter 3.2).

Usage scenarios for matching object characteristics

Input

Two datasets at the micro level. The data may be administrative data or statistical data.

Output

One integrated micro-dataset.

Methodological options

Option 1: Unweighted matching of object characteristics.

Option 2: Weighted matching of object characteristics.

Option 3: Record linkage.

Evaluation of the options

The following diagram shows the decision tree for application of the methods:

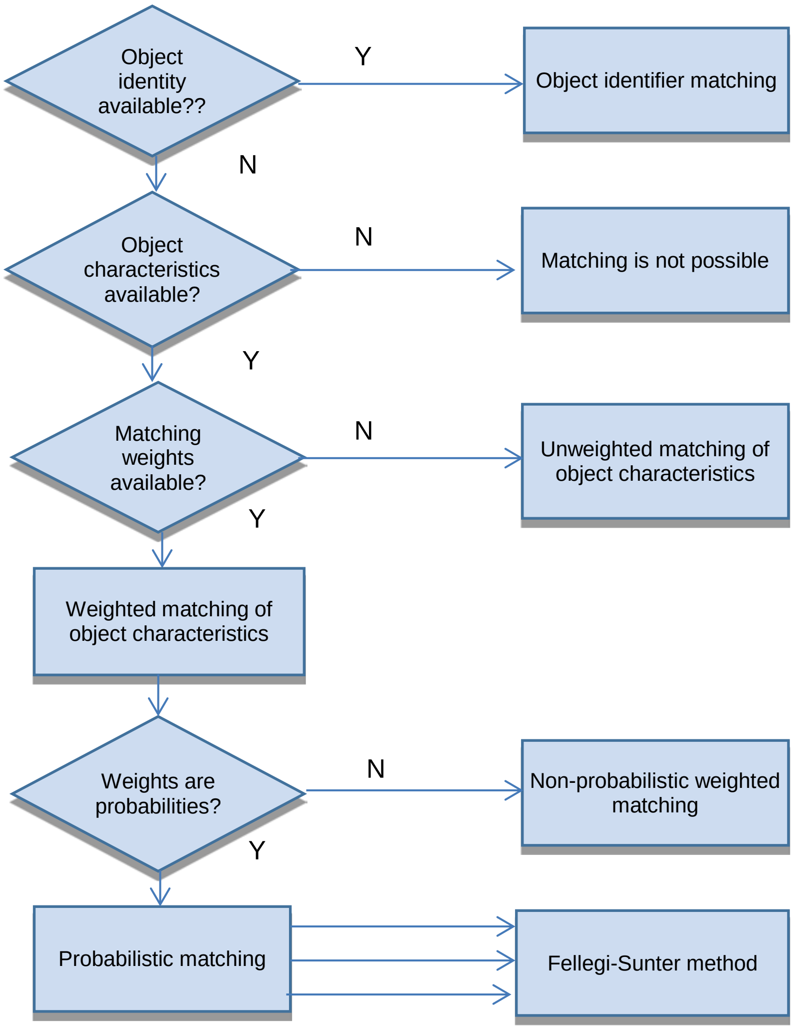
Object identity available??
Y
Object identifier matching
N
Object characteristics available?
N
Matching is not possible
Y
Matching weights available?
N
Unweighted matching of object characteristics
Y
Weighted matching of object characteristics
Weights are probabilities?
N
Non-probabilistic weighted matching
Y
Probabilistic matching
Fellegi-Sunter method

Figure 1: Decision tree for application of the methods

Examples

Several example application areas where record linkage is an important component of larger information systems, of government and business processes, or of research endeavours: National census; Health sector; National security; Crime and fraud detection and prevention; Business mailing lists.

Statistics Canada has one of the most longstanding and extensive expertise in record linkage. Internationally, Statistics Canada is a leader in developing linkage methods, including the theoretical work which Fellegi and Sunter (1969) began. Today, many statistical agencies throughout the world are inspired by Statistics Canada's model and practices in record linkage. Statistics Canada undertook an ambitious project on data integration in the business sector, with its initial phase beginning in 2008. This project, known as the Linkable File Environment (LFE), was implemented through the creation of a relational database that associates numerous information sources (surveys and administrative data) with the Business Register (BR), which constitutes the reference database.

New Zealand has considerable experience with record linkage. Since 1997, this has included the following projects: Linked Employer-Employee Data (LEED); the Longitudinal Business Database (LBD) prototype; the Household Labour Force Survey (HLFS); Employment Outcomes of Tertiary Education (EOTE) and the Student Loans and Allowances (SLA) integrated dataset.

Like some other countries, Germany has a private research centre entirely devoted to data integration services, the German Record Linkage Centre (German RLC). This centre is currently conducting some 15 record linkage projects. One example concerns the integration of data from the German Business Register (URS) with administrative data from institutions such as the Federal Employment Agency and the Deutsche Bundesbank. This program is called the KombiFi (Kombinierte Firmendaten für Deutschland).

References

- Harron, K., Goldstein, H. & Dibben C. (2015). *Methodological Developments in Data Linkage*. John Wiley & Sons, Ltd (Wiley Series in Probability and Statistics), 259
- Ivan P. Fellegi & Alan B. Sunter (1969) *A Theory for Record Linkage*, Journal of the American Statistical Association, 64:328, 1183-1210,
- Memobust Handbook: Module Micro-Fusion.
- The AdminData ESSnet: Use of Administrative and Accounts Data for Business Statistics, WP3 - Methods of Estimation for VariablesESSnet ISAD project, WP 2 Recommendations on the use of methodologies for the integration of surveys and administrative data.
- ESSnet Data Integration, WP 1: State-of-the-art.
- P. Christen (2012). *Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and duplicate Detection*. Springer, Berlin.
- A. Wallgren, B. Wallgren (2014). *Register-based Statistics Statistical Methods for Administrative Data.* John Wiley & Sons, Ltd.

6.2 Methods based on statistical matching (data fusion)

Description of the problem

Statistical matching (also named data fusion or synthetic matching) refers to a series of methods whose objective is the integration of two (or more) data sources referring to the same target population. The data sources are characterised by the fact they all share a subset of variables (common variables) and, at the same time, each source observes distinctly other subsets of variables. Moreover, there is a negligible chance that data in different sources observe the same units (disjoint sets of units).

In statistical matching, there are two non-overlapping sources, in the sense that the two sets of units collected in the two surveys are distinct. They refer to the same target population, and the variables of interest for the statistical analyses are available distinctly in the two surveys. Due to the nature of the datasets, it is not possible to create joint information on these variables by means of their common identifiers.

Generally speaking, this problem has been considered as an imputation problem. Taking two files, one file, e.g. A, being considered the recipient and the other the donor file, e.g. Z, the statistical matching procedure consists in imputing Z in A by means of the available common information X.

Methods and workflow for statistical matching

Two main approaches can be delineated in terms of outputs that can be obtained through matching:

The macro approach refers to the identification of any structure that describes relationships among the variables not jointly observed of the datasets, such as joint distributions, marginal distributions or correlation matrices.

The micro approach refers to the creation of a complete micro-data file where data on all the variables is available for every unit. This is achieved by means of the generation of a new dataset from two datasets based on an informative set of common variables. The result is called a dataset of ‘synthetic micro records’.

The synthetic dataset is the basis of further statistical analysis, e.g. microsimulations. The word synthetic refers to the fact that the records are obtained by integrating the available datasets rather than direct observation of all the variables.

Usually the matching is based on the information (variables) common to the available data sources and, when available, on some auxiliary information (a data source containing all the interesting variables or an estimate of a correlation matrix, contingency table, etc.). When the additional information is not available and the matching is performed on the variables shared by the starting data sources, then the results will rely on the assumption of independence among variables not jointly observed given the shared ones.

Methods for statistical matching

The following three classes of methods can be used for statistical matching:

- Parametric approach. A model characterised by a finite number of parameters is explicitly considered; once its parameters are estimated it is possible to impute the values of the missing variables via conditional expectation (conditional mean matching) or by drawing values from the predicted distribution.
- Nonparametric approach. Does not require the explicit usage of a model and is more flexible in handling complex situations (a lot of variables of mixed type, categorical and continuous). The most used nonparametric techniques in statistical matching derive from hot deck methods applied in sample surveys to fill in missing values. Usually the objective is to create the synthetic dataset by imputing the missing variables in the recipient dataset. Imputed values are those observed in a similar statistical unit observed in the donor dataset. Three hot deck methods have been used in statistical matching:

  a) random hot deck - the record is taken at random from the ones observed in the donor dataset;

  b) rank hot deck - the records in the recipient and donor files are ordered according to one common variable and the corresponding cumulative distributions functions are computed;

the donor is chosen among those whose value of the cumulative distribution function is the nearest to the corresponding value in the recipient;

c) distance hot deck - taking a distance function of the common variables observed in the two files, each record in the recipient is imputed with that record in the donor whose observed common variables are the nearest to the ones observed in the recipient record.

- Mixed approach. This class of techniques mixes parametric and nonparametric approaches. More precisely, initially a parametric model is adopted, and then a completed synthetic data set is obtained by means of some hot deck procedures. An example is the predictive mean matching method (Little, 1988)

Workflow for statistical matching

The workflow is a specification of GSBPM sub-process 5.1: “Integrate data” for this specific usage scenario.

Two (or more) datasets containing micro-data, whose observed sets of units are disjoint, and the observed variables admit a strict subset of common variables.

Statistical matching techniques are usually applied to investigate the relationship between two variables, Y and Z, never jointly observed in the available data sources, by considering the available common information, usually X variables.

Remarks on the workflow

When no auxiliary information is available the statistical matching is based on an assumption, such as conditional independence. These assumptions cannot be verified on the available data. If the analyst does not consider it valid, the statistical matching cannot be performed.

Usage scenarios for statistical matching

Input

Non-overlapping micro-data sources without coverage problems; some of the variables are overlapping in both sources, some of the variables are only in single sources.

Micro-data covering the target population, with overlapping units between the different sources, and without coverage problems.

Output

A complete dataset for the joint variables.

In the macro approach, structural characteristics of the joint distributions.

Methodological options

Option 1: Parametric approach.

Option 2: Nonparametric approach.

Option 3: Mixed approach.

Evaluation of the options

- When the output is essentially micro and many observations are available, a **nonparametric approach** (option 2) can be preferable.
- A **parametric approach** (option 1) is generally preferable when the objective of the statistical matching is macro. When a parameter/indicator should be estimated (for instance correlation) and not many observations are available, the model introduced in the estimation may be more efficient since it is generally more parsimonious than nonparametric approaches.

- A mixed approach (option 2) is generally adopted when the objective is micro and it is difficult to introduce an explicit model. In fact, the main reason for the introduction of this approach is that it exploits the advantage of a model (being more parsimonious for the estimation) but providing final 'live' data, i.e. really observed (synthetic data set production). Moreover, it may be useful in those situations when the model assumptions do not hold exactly.

Remarks on methods

The most common approaches are nonparametric and belong to the hot deck imputation methods:

The random choice of random hot deck is often done within groups obtained by considering subsets of homogeneous units characterised by presenting the same values for one or more common variables X (usually categorical).

Distance hot deck is widely used in the case of continuous variables. The donor unit is the closest to the given recipient units in terms of a distance measured by considering all or a subset of the common variables X. In general the methods based on distances pose the problem of deciding the subset of the common variables X to be used for computing it. Using all or too many common variables may affect negatively the matching results because variables with low predictive power on the target variable may influence negatively the distances.

The rank hot deck distance method searches for the closest donor for the given recipient record with distance computed on the percentage points of the empirical cumulative distribution function of the (continuous) common variable X being considered. Considering the percentage points of the empirical cumulative distribution provides values uniformly distributed in the interval [0, 1]; moreover, this allows to compare observations when the values of X cannot be directly compared because of measurement errors which however do not affect the "position" of a unit in the whole distribution.

Parametric methods have also been proposed (e.g. through imputations by means of regression functions, with the possibility to add noise around the function itself), despite their dependency on the normality assumption being very strong. For this reason imputations by parametric models are also mixed with non-parametric methods, leading to the so-called statistical matching mixed methods.

Examples

Eurostat developed a pilot study focused on testing the feasibility of using matching techniques in order to obtain joint distributions for various dimensions of quality of life, drawing on variables collected through two main sources: the European Union Statistics on Income and Living Conditions (EU-SILC) and the European Quality of Life Survey (EQLS).

In Italy, the joint analysis of household income and expenditures can be obtained by the joint use of two samples, the Household Budget Survey (HBS, managed by Istat) and the Survey on Household Income and Wealth (SHIW, managed by the Bank of Italy).

Statistics Canada imputes the Survey of Labour and Income Dynamics (recipient) with three other datasets: a sample of anonymized Personal income tax return data, a sample of Employment Insurance claim histories, and the Survey of household spending.

References

- Donatiello D., D’Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2016) *The role of the conditional independence assumption in statistically matching income and consumption*. Statistical Journal of the IAOS, 32, 667-675, DOI 10.3233/SJI-161000.
- D’Orazio M., Di Zio, M., Scanu M. (2006) *Statistical Matching – Theory and Practice*. John Wiley & Sons Ltd.
- Little R.J.A. (1988) “Missing-data adjustments in large surveys.” Journal of Business & Economic Statistics 6: 287-296.

7 Estimation methods for macro-integration

The presentation of tables is one of the most important tasks in official statistics. Such tabular information about the economic, socio-economic and socio-demographic facts has to be understood not as standalone information but has to give a coherent picture for the population. Macro-integration is an activity which aims at the integration of different data sources in such a way that the information is coherent and follows a so-called one figure policy or univalency. This means that the information about a certain characteristic is the same in all tables.

Two approaches are possible for achieving this goal. The first one is the more traditional approach based on methods for adjustments of the macro-data without taking into consideration the underlying micro-data. The other one is an approach which adapts the underlying micro-data in such a way that coherence between different tables is achieved.

The methods are not specific to administrative data but can be used for production of any kind of multisource statistics.

7.1 Macro-integration based on adjustment of macro-data

Description of the problem

Macro-integration is the process of integration of data at the aggregated level. Various types of errors can cause discrepancies between two or more independent macro-data sources. In the case of survey data, the most important errors are errors due to measurement of the variables (measurement error, processing error) and errors due to the representation of the population (sampling error, non-response error and correction error). One can identify similar measurement errors in the case of administrative data and for the representation of the population one can distinguish coverage errors, linking errors, and correction errors. Consequently, the combination of different sources in one table will lead to inconsistency of the figures. Macro-integration aims at the reconciliation of such inconsistencies at the macro level. Contrary to univalent estimation the adjustment is done only for the aggregated data but leaves the underlying micro-data unchanged. Depending on the complexity of the combination of the underlying problem, different methods can be used for adjustment. Hence macro-integration is an activity of the sub-process 6.2 of the GSBPM.

Methods and workflow for macro-integration

Methods for macro-integration

Three basic methods for macro-integration will be considered:

- *Manual integration*. This method is also known under the heading “balancing”. The aim is to obtain consistent tables for which a balance equation between the variables is defined. Usually the finding of inconsistencies in such balance equations is done manually and the reconciliation of the inconsistencies is done by using rules which are based on understanding the possible reasons for inconsistencies. A typical example of finding an explanation for the inconsistencies is comparison of the results in the previous time period.

- *RAS*. The RAS method is a well-known and widely used method for data reconciliation. It is also known under the name "Iterative Proportional Fitting" in the analysis of contingency tables. RAS is also called "matrix raking" or "matrix scaling" in computer science. Its aim is to achieve consistency between the entries of some non-negative matrix and pre-specified column totals and row totals. It is very easy to apply and to understand. However RAS has a narrow scope of applicability, for example, it can only be applied to non-negative matrices. RAS is a recommended method if the data are presented as a two-dimensional table and the rows and columns of the data are measured with higher precision than the cells. A drawback of RAS is, as with other methods for macro-integration, that there is no direct connection between the data in the table and the underlying micro-data. There exist also some generalizations of RAS for matrices with negative entries and for fixing, not only the row and column totals, but also for an arbitrary subset of the matrix elements. The RAS method is described in detail in the Memobust Handbook: Module Macro-Integration – Method: RAS (2014).
- *Stone's method*. This method adjusts macro-data for which a number of linear constraints are defined. It is an iterative method which solves a weighted quadratic optimization problem. Besides the data itself, knowledge about the reliability of the initial data is necessary, in particular the covariance matrix of the data. If the covariance matrix is not known, estimates for the relative variance can be used. Stone's method can be applied to more complex structures of macro-data, for example a set of tables.

General workflow for macro-integration

Macro-integration is done mainly in GSBPM sub process 6.2: "Validate outputs" but also in sub-process 5.7: "Calculate aggregates".

Macro-integration is done in two stages:

- In the first stage the data sources are adapted to comply with the correct definitions;
- In the second stage, which is called data reconciliation, discrepancies at the aggregated level are resolved. At first an adjustment for major errors is done, usually by manual integration. In the next step adjustment for the noise is done.

Remarks on the workflow for macro-integration

In the simplest case of macro-integration the goal is the estimation of a table defined by two variables for which the univariate distribution of the variables is known. The origin of the variables may be survey data or administrative data.

More complex specifications, like considerations of many different tables, are possible.

A specific situation is the integration of tables for which balance equations are defined, for example supply and use tables in economic statistics.

Usage scenario for macro-integration

Input

Two or more macro-datasets which may be administrative data or statistical data. Probably some constraints for the figures in the table are defined.

Output

One integrated macro-dataset.

Methodological options

Option 1: Manual integration.

Option 2: RAS.

Option 3: Stone's method.

Evaluation of the options

Due to the fact that the methods are designed for specific macro-data configurations, the decision about the method can be formulated as follows:

- If the macro-data have to fulfil balance equations, these equations can be broken down to different cells of the tables, and there is knowledge about the most plausible reasons for inconsistencies, then apply manual integration.
- If the macro-data are given as a rectangular table and the reliability of the row and column totals is higher than the reliability of the cells, use RAS.
- If the macro-data are more complex structured (for example a set of tables), use Stone's method.

Examples

The Module Macro-integration in the Memobust handbook describes examples from business statistics for each method.

References

- Memobust Handbook: Module Macro-Integration.
- Eurostat Manual of Supply, Use and Input-Output Tables, Eurostat, European Commission, 2008, Chapter 14: Updating and projecting input-output tables.

7.2 Macro-integration based on adjustment of micro-data

Description of the problem

The goal of univalent estimation is a realisation of the so-called one figure policy. This means that estimates for the same phenomenon in different tables should be the same even if the estimates are based on different underlying data sources. This principle is called univalency. Hence the goal is rather similar to macro-integration but the methods for obtaining this goal are quite different. In macro-integration the original micro-data are not changed, in the case of univalent estimation the underlying micro-data are modified.

In the case of using survey data and administrative data the estimates for the same phenomenon will be different even when there are no errors in the data. The general philosophy of all univalent estimates is to modify the underlying micro-data and use for the production of the macro-data the modified micro-data. This can be done in a repeated way for different tables or in a one-step update of the micro-data.

Methods and workflow for macro-integration based on adjustment of micro-data

Methods for macro-integration based on adjustment of micro-data

Three methods are considered:

- Repeated weighting. In this approach population tables are estimated sequentially and each table is estimated using as many sample units as possible in order to keep the sample variance as low as possible. The combined data from the data sources are divided into rectangular data blocks. Such a block consists of a maximal set of variables for which data on the same units has been collected. The data blocks are chosen such that each table to be estimated is covered by at least one data block.
- Mass imputation. This method imputes all population units for all variables for which no values were observed a value. In this way a complete rectangular dataset for the whole population is obtained. All macro-data are produced from this complete dataset.
- Repeated imputation. Repeated imputation is an iterative approach similar to repeated weighting. For the different tables separate imputations are assigned for the units and the estimation of the tables are obtained step by step. If estimates for a variable have already been produced in a previous table, the variable is calibrated.

General workflow for macro-integration based on adjustment of micro-data

Depending on the methods used for univalent estimation, the GSBPM sub-processes 5.4: “Edit & impute”, 5.6: “Calculate weights”, 5.7: “Calculate aggregates” and 6.2: “Validate outputs” are relevant.

- Descriptive analysis of the underlying micro-data;
- Decision about the intended tables;
- Decision about the algorithm for repeated weighting, or about the used imputation methods and variables for imputation;
- Computation of the weights or the imputed values;
- Analysis of the results;
- Documentation of the process.

Remarks on the workflow

Data from a data source covering the entire population can simply be counted. Data only available from surveys are weighted by means of regression weighting. In that case starting weights need to be assigned to all units in the block to be weighted. For a survey one usually starts with the inverse inclusion probabilities of the sample units, corrected for response selectivity. For a data block containing the overlap of two surveys, one usually begins with the product of the standard survey weights from each of the surveys as starting weight for each observed unit.

When estimating a new table, all cell values and margins of this table that are known or have already been estimated for previous tables are kept fixed to these known or previously estimated values. This is achieved by using regression weighting, where the starting weights are adjusted by calibrating to known or previously estimated values. This ensures univalency of the cell values and margins of the new table and previous estimates.

The method relies on the used imputation model, which must capture all relevant variables and all relevant relations between the variables. Furthermore, the estimates for the parameters in the imputation model must be accurate. With respect to possible relations one must take care of structural edits. Documentation of the method should include the specification of the imputation model.

There are different proposals for the computation of the weights in repeated weighting. The traditional method has the drawback that in some cases finding a consistent solution is impossible. Furthermore, the solution is often only a sub-optimal solution and depends on the weighting sequence. Recently, a branch and bound algorithm for overcoming such problems was proposed.

In the case of mass imputation the results can be used only for the intended purpose. For example, if a relation between a variable A and a pair of variables (B, C) is part of the imputation model, then interpretation of the relation between variables A and B is not possible.

Usage scenarios for macro-integration based on adjustment of micro-data

Input

Micro-datasets.

Output

Integrated micro-dataset.

Methodological options

Option 1: Repeated weighting

Option 2: Mass imputation

Option 3: Repeated imputation

Evaluation of the options

- In general, the application of mass imputation has some advantages. Computation can be done by standard software and evaluation is possible using the evaluation by training and test data.
- Repeated mass imputation has similar problems like repeated weighting.
- Repeated weighting is from the computational point of view the most challenging method and needs extensive testing about the sequence of the tables. If possible the new branch and bound algorithm should be used.

Examples

Simple demonstration examples can be found in the documentations of the methods.

References

- ESS Vision 2020 ADMIN WP2, Statistical methods, Task 5 Summary Sheets of methods: 2f Repeated weighting.
- ESS Vision 2020 ADMIN WP2, Statistical methods, Task 5 Summary Sheets of methods: 2g Mass imputation.
- ESS Vision 2020 ADMIN WP2, Statistical methods, Task 5 Summary Sheets of methods: 2h repeated imputation.
- Daalmans, J. (2015): *Divide and Conquer solutions for estimating large consistent table sets*
- Knottnerus, P., & van Duin, C. (2006). *Variances in repeated weighting with an application to the Dutch Labour Force Survey*. Journal of Official Statistics, Stockholm, 22(3), 565
- Jacco Daalmans, (2017) *Mass Imputation for Census Estimation*, Statistics Netherlands, The Hague/Heerlen/Bonaire 2017

8 Using administrative data for creation and maintenance of registers

Statistical registers are of utmost importance for statistical data processing. Typically, registers play the role of a data coordination tool, and integrate data from several sources, both statistical and administrative. This may be done by linking records by means of common identifiers, or by using matching techniques. A characteristic of a statistical register is that it is updated periodically, e.g. annually.

Using only statistical data sources for the creation of a statistical register is usually not feasible due the high costs. Given the backbone role statistical registers play, in many countries legal acts exist for the registers, for example business registers. The creation and maintenance of the statistical registers rely to a great extent on administrative data, in particular on administrative registers.

This chapter considers the role of administrative data in two important statistical processes: the creation of a register using administrative data and the update of statistical registers. The guideline for creation of a statistical register is defined by GSBPM adapted to this specific application. Different methods for the sub-processes can be used. These methods define the options for register production. Of utmost importance are the sub-processes integration and alignment of units and measurements considered in chapters 4 and 5.

8.1 Creation of statistical registers from administrative registers

Description of the problem

A statistical register is a register created and maintained for statistical purposes according to statistical concepts and definitions, and under the control of statisticians. It is typically created by transforming data from registers and/or other administrative data sources.

Among the many functions of statistical registers, the following are of particular relevance: (i) Identification and construction of statistical units; (ii) Unambiguous documentation of the links between units; (iii) Evidence on populations; (iv) Evidence on the main characteristics of the units; (v) Documentation of the classification of units; (vi) Building a sample frame, i.e. establishing a common framework for organising and coordinating statistical surveys by providing a harmonised sampling frame according to stratification characteristics; (vii) Providing the basis for the grossing-up of results derived from sample surveys; (viii) Keeping track of creations and closures, and other demographic events of units; (ix) To be a statistical output by itself, to produce structural information on the specific target population (persons, households, enterprises, farms, etc.).

Methods and workflow for statistical registers

Methods for creation of a statistical register

Creation of a statistical register requires different production methods. The most important are the following:

- Conceptual data modelling for defining the structure of the register;
- A model which supports the storage of the history of the statistical units in the register;
- Selection of appropriate statistical and administrative data sources for the register;
- Organisation of the data exchange from the administrative owners and definition of an update policy;
- Evaluation of the quality of the different sources;
- Methods for alignment of units and measurements;
- Methods for data integration;
- Methods for editing and imputation;
- Methods for estimation of variables;
- Methods for ensuring privacy and security.

General workflow for the creation of a statistical register

The creation of a statistical register resembles the GSBPM adapted to the specificities of the register. The following steps of the GSBPM are of utmost importance:

- Identify the needs of the statistical register;
- Identification of target population;
- Identification concepts for the statistical units and the variables;
- Design the data collection from administrative sources;
- Collect the data from the administrative and statistical sources;
- Run the data collection;
- Align the different units and measurements in the sources to the statistical concepts;
- Integrate the data;
- Edit and impute the data;
- Derive new variables and units;
- Finalise the statistical register;
- Document the processes used for creation of the register;
- Document the output quality of the register.

Remarks on the workflow for the creation of statistical registers

It should be noted that in some cases the variables in administrative registers are based on legal acts, hence they cannot be changed by the statistical institution. Differences between administrative units and statistical units and between administrative variables and statistical variables must take this into account in the development of methods for alignment of units and measurements. For example, differences in the definition of units and different classifications for variables.

Using administrative registers for the creation of statistical registers is facilitated if the administrative registers are maintained by the NSI.

An important issue is the identification of the units by key variables.

Usage scenarios for the creation of statistical registers

Input

Statistical micro-data, administrative micro-data (in most cases administrative registers); the data must cover the statistical population of interest; redundancy of variables can be helpful for the detection of possible errors.

Output

A statistical register which can be used by the NSI for different purposes, outlined in section 8.1.1.

Methodological options

Option 1: Use statistical and administrative sources which have some overlap in the variables and in the units. The quality of the sources is known and the register update policy of the source owner is transparent.

Option 2: Use disjoint sources of good quality with transparent update policies and temporal stability.

Option 3: Use administrative data which are not well documented and with low stability.

Evaluation of the options

- The best option is option 1;
- Option 2 is acceptable;
- Option 3 should be avoided.

In all cases, the documentation of the input quality, the process quality and the output quality should be provided.

Examples

In Austria, the statistical business register (URS) and the Austrian administrative business register (URV) are both maintained by Statistics Austria, a good practice that facilitates the setup and the maintenance of the URS. Besides the administrative register URV, a number of other sources are used for the register, in particular: data sources from the Austrian Chamber of Commerce, the Austrian tax information system, the register of the Main Association of Austrian Social Security Institutions, data from external registers such as information about educational institutions. Statistics Austria has defined a comprehensive list of possible sources and workflows for the integration of the various sources into the URS. For all data providers a time schedule for data provision is defined. Most of the data are processed automatically and data matching uses advanced text processing methods for identification of keys. Manual editing is minimized by editing rules which are regularly updated. Of utmost importance is the classification of enterprises according to NACE, Rev.2. An interface for enterprises facilitates the contacts with the business owners about the classification.

In Italy, the statistical business register (SBR) is maintained by the NSI (Istat) and the administrative business register is maintained by the Chamber of Commerce. To build up the last version of the SBR, Istat used the 2011 census survey on economic units and a large number of administrative sources, the main ones coming from the Chamber of Commerce, tax authorities and social security institutions. The data are processed automatically. Manual editing is minimized by editing rules which are regularly updated.

The 2011 economic census was the last economic census in Italy. It was done for two main reasons: (i) to check the quality of the administrative sources used to build the SBR, (ii) to check the quality of the methods developed by Istat to align the administrative units and variables to the statistical concepts.

Istat produces a declaration of SBR quality. The declaration includes a set of indicators measuring the quality of various components of the register. The main components are: timeliness, coverage, completeness and accuracy.

References

- Quality of registers: ESS Vision 2020 ADMIN project (KOMUSO, WP 2)
- Li-Chun Zhang, John Dunne, Trimmed dual system estimation, Capture-Recapture Methods for the Social and Medical Sciences, (2017) Chapman and Hall, New York
- Guideline on Statistical Business Registers, UNECE 2015
- A. Wallgren, B. Wallgren (2014). *Register-based Statistics Statistical Methods for Administrative Data*. John Wiley & Sons, Ltd.

8.2 Updating of statistical registers using administrative registers

Description of the problem

Registers should give accurate and timely information about the population defined by the statistical units in the statistical register. Hence the register must be updated periodically[5]. This update should reflect the dynamics of the underlying population. Typical events for the statistical units in the register are the creation of new statistical units (birth), the end of the activity of the register unit (death), the merging of register units, and the splitting of register units. These changes must be considered for all unit types (persons, household, enterprise, farms, etc.).

Methods and workflow for the updating of statistical registers

Methods for the updating of statistical registers

The following methods are used to update a register:

- Workflow for tracking the update process of the statistical units;
- Identification of organisational and legal changes in administrative sources from “t” to “t+1”;
- Methods for integration of data sources;

[5] The timetable to update the register (from “t” to “t+1”) can be monthly, quarterly, yearly, or every “n” years.

- Comparison of the statistical units in two micro-datasets in “t” and “t+1”;
- Comparison of the value domains of statistical variables in “t” and “t+1”.

General workflow for updating statistical registers

The workflow resembles the different GSBPM phases for statistical production. The most important are the following topics:

- Creation of an update schedule for new information from the different providers;
- Integration of the existing statistical registers;
- Identification and resolution of the mismatches in the statistical units, between “t” and “t+1”;
- Identification and resolution of the mismatches in the statistical variables, between “t” and “t+1”;
- Decision rules on how to use the new information from the different sources;
- Creation of a new version of the statistical register;
- Documentation of the changes.

Remarks on the workflow

The resolution depends on the quality of the new information in the administrative sources. As regards the existence of units, in the ideal case the information about the units’ dynamics is documented in more than one register. The sign of life defined in accordance to statistical concepts is often a useful strategy.

With respect to the variables, the decision about an update depends on the quality of the variables in the administrative sources. The ideal case when updating a value is to have several sources for the same variable. If necessary, a useful strategy is to have a survey (current or specific) for checking the quality of variables.

An important issue is consideration of administrative delay, which means that it usually takes some time for changes in the units to be documented in the register.

Usage scenarios for updating statistical registers

Input

A statistical register together with actual statistical sources (surveys) and administrative sources.

Output

An updated version of the statistical register.

Methodological options

Option 1: For the update, use overlapping administrative sources that are of good quality and have no changes in their legal and organisational aspects.

Option 2: For the update, use sources with changes in legal and organisational aspects.

Option 3: Use new administrative sources without any checks on the quality.

Evaluation of the options

- Option 1 is the best alternative;
- Option 2 is acceptable;
- Option 3 should be avoided.

In all cases, it is important to document the processes.

Examples

In the Austrian business register updating and maintenance of the register is done with different periodicities depending on the data sources. The basic information about the enterprises is updated daily, whereas information from the tax register and the Main Association of Austrian Social Security Institutions monthly. The main tasks in the maintenance are: capturing of new units, changes in the structure of units such as new local units, new accounts, change in classification (NACE, Rev.2), actualisation of turnover and number of employees, changes in the activity status (close down). For all types of changes relevant data sources are identified. The fact that the administrative register (URV) and the statistical register (URS) are maintained inside the NSI facilitates coherence between the two information systems. For the update, processes workflows are defined which support tracking of the update process. The register stores the history of all units inside the register.

In Italy the SBR is updated yearly. The main task in the maintenance are: capturing of new units (real birth), changes in the structure of units such as new local units, change in classification (NACE, Rev.2), actualisation of turnover and number of employees, changes in the activity status (real death). The basic sources are the same every year (the Chamber of Commerce, tax authorities, social security institutions), but new sources are added to improve the quality of the output. To check the quality of the SBR over time, Istat has defined a special sample survey. This annual survey is used to verify over time the quality of administrative sources in updating variables and units and the quality of the statistical methods developed to align these to the statistical concepts. Every year checks for specific sub-populations are defined. Changes in the administrative sources (legal changes on target population, units and variables and changes in the administrative process) are checked by the special survey and by formal agreements defined between Istat and the data owner. The quality indicators described in section 8.1.4 are updated yearly.

References

- Guideline on Statistical Business Registers, UNECE 2015
- A. Wallgren, B. Wallgren (2014). *Register-based Statistics Statistical Methods for Administrative Data*. John Wiley & Sons, Ltd.

9 Direct usage of administrative data in statistical processing

This chapter considers the use of administrative data for obtaining statistical outputs. In this application the administrative data play a leading role. The methods used are based on the application of the GSBPM and possible options are defined in many ways: the sequencing of the different steps, the options discussed in the chapters about the GSBPM sub-processes, and options defined by organisational and administrative constraints of the NSI. Hence, the definition of a specific workflow has many decision points for preferred options.

Two different scenarios are considered. The first one is so-called direct tabulation where administrative data are used as the only source for statistical production. The first one is a more direct use of an administrative source for tabulation while the second one is combining different sources, administrative and statistical, in the production of statistical products.

9.1 Combining administrative sources for tabulation of statistical results

Description of the problem

The goal of this application is the production, from one or more administrative sources, of a statistical data matrix where the rows represent the statistical units of the population and the columns the statistical variables of interest. From this data matrix the requested tables are produced. An important issue is the definition of the reference population for identification of the missing statistical units and hence, eventually, how to treat them. The methods are similar to those used for the production of a statistical register considered in Chapter 8.

Methods and workflow for direct tabulation

Methods for combining administrative sources for tabulation

The methods are virtually the same as for the creation of registers:

- Conceptual data modelling for defining the structure of the data corpus from which the requested table can be retrieved;
- Analysis of metadata;
- Methods for data retrieval in a relational database;
- Selection of appropriate administrative data sources for the requested table;
- Organisation of the data exchange from the administrative sources;
- Evaluation of the quality of the different sources;
- Methods for alignment of units and measurements;

  - Methods for data integration;
  - Methods for editing and imputation;
  - Methods for estimation of variables;
  - Methods for validation of variables;
  - Methods for ensuring privacy and security.

General workflow for combining administrative sources

Combining administrative sources for direct tabulation follow the GSBPM. The following steps of the GSBPM are of utmost importance:

- Identification of administrative source/sources according to the needed information;
- Collection of the metadata and validation of the source/sources with respect to the target contents;
- Design the data collection from administrative sources;
- Implementation of data collection tools;
- Collection of the data from the administrative sources;
- Validation of the data with respect to coverage;
- Align the different units and measurements in the sources;
- Integrate the data;
- Edit and impute the data;
- Derive new variables and units;
- Building a data corpus for the administrative data under consideration;
- Produce the intermediate data matrix;
- Calculate aggregates and prepare the requested table;
- Use disclosure control;
- Document the processes used for creation of the statistical table.

Remarks on the general workflow

The different sources are linked with the list that represents the target population in order to identify duplications, missing units and the units not included in the list by definition. Duplicated units and units not in the target population are deleted.

To join together the different sources the record linkage techniques (unweighted matching, weighted matching and probabilistic record linkage) are used.

The integration and the successive steps of data treatment can be developed in a different manner as described in the options for data editing and imputation.

The integrated data can contain missing units, missing items, inconsistencies and conflicting quantitative variables. The scope of filling in all the units belonging to the population can be reached treating the missing units with mass or donor imputation techniques based on the variables available in the list.

A workflow for sequencing editing and imputation of the variables should be formulated together with the necessary methods.

Data treatment and data integration can be used in different sequences. Two basic solutions can be applied: treatment before integration of the sources, data integration and afterwards data treatment for the integrated dataset; alternatively, the single sources are treated and then they are integrated (see also chapter 3 about editing and imputation). Generally the first solution is preferred because more complete information for aligning, editing, imputing and deriving new variables can be used, even if the editing and imputation procedures can be more complex. Moreover, in the second sequence, a final step of editing and imputation on the integrated data is needed because the treatment of the single source does not assure the consistency of the final dataset. An intermediate solution is to limit the single sources to a “light” treatment and then to join them and apply the “complete” treatment on the integrated data.

Usage scenarios for combining administrative sources

Input

Administrative micro-data sources from different data owners.

Output

A statistical product.

Methodological options

Methodological options can be defined for the different sub-processes. For the most important steps these options were formulated and discussed in the previous chapters. Besides these options the following options for the division of the workflow inside the NSI can be formulated:

Option 1: Decentralised processing. Inside the NSIs the procedures described above may be organised in a different way. For example, a centralized input point that distributes the administrative sources from different input points related to the content of administrative sources or a combination of centralized/decentralized input points. Depending on the organisation, the responsibility of the steps needed to transform the administrative data into statistical data, and hence the related tasks, can be distributed differently according to the knowledge and skills of the different departments.

Option 2: Centralised processing. A centralised organisation avoids redundancy and can better implement methodological innovation. On the contrary, decentralized input points can better utilise the available resources and assure higher overall quality in using the different expertise of the staff.

Option 3: Mixed processing. The combination of a centralised and a decentralised organisation can balance the pros and cons of the two kinds of organisation. For example, a centralised point can manage all the tasks related to the relationships with the owners, the data and metadata collection and storing of the data. Also a first data treatment for identifying and correcting duplications and for standardising the input, as well as the alignment of units and keys for common variables can be centralised. The more detailed analysis, cleansing and harmonisation of the content variables are decentralised to the experts.

Evaluation of the options

- From the administrative options, Option 3 (mixed approach) is preferable;
- Option 2 (centralized approach) is acceptable;
- Option 1 (decentralised processing) is not recommended.

However, the choice of options must be seen in connection with the organisational and administrative background of the office.

For evaluation of the production, the criteria defined in the KOMUSO project about output quality should be used.

Examples

The application of this scenario for the register based census in Austria in 2011 is documented briefly in Deliverable D3 of the project Good practices when combining some selected administrative sources of WP 2, Task 2.2 of the ESS.VIP ADMIN project.

A further example is the production in Italy in the context of Structural Business Statistics, using all the available administrative sources of a "Frame" containing the most important economic variables for the entire business population, from which tables and indicators are calculated (see the reference).

References

- Quality of multisource statistics: ESS Vision 2020 ADMIN project (KOMUSO, WP 2).
- A. Wallgren, B. Wallgren (2014). *Register-based Statistics Statistical Methods for Administrative Data*. John Wiley & Sons, Ltd.
- O. Luzi, R. Monducci (2016). *The new statistical register "Frame SBS": overview and perspectives*. Istat Rivista di Statistica Ufficiale n. 1 – 2016.

9.2 Substitution and supplementation for direct collection

Description of the problem

In this case, administrative sources are also used to produce statistics but they are not sufficient to cover all the informative needs and hence they are used in combination with statistical data. Two cases can be distinguished:

- Split population approach. The population is divided essentially into two non-overlapping groups; for the first group reliable administrative data are available and for the second group data from a census or from one or more statistical surveys or registers are available.
- Split data approach. For some of the target variables, administrative data are available for the entire target population but the data do not cover all the needed information. This missing information can be obtained from one or more statistical surveys or registers.

The discriminant point here, with respect to the combination of only administrative sources in section 9.1, is that sub-processes concerning sampling, in particular the calculation of weights, must be considered. Furthermore, the variables in the administrative source can be part of the survey and are available for the sample used in the survey.

Methods and workflow for substitution and supplementation

Methods for substitution and supplementation

In this case, the most important methods for the production of the statistical outputs are:

- Selection of appropriate statistical and administrative data sources;
- Analysis of metadata;
- Organisation of the data exchange from the administrative sources;
- Evaluation of the quality of the administrative sources;
- Methods for alignment of units and measurements;
- Methods for data integration;
- Methods for editing and imputation;

  - Methods for estimation of variables;
  - Methods for validation of variables;
  - Methods for calculation of weights;
  - Methods for calculating aggregates.

General workflow for substitution and supplementation

Substitution and supplementation for direct collection follow the GSBPM. The following steps of the GSBPM are of utmost importance:

- Collection of the metadata and validation of the source/sources with respect to the target contents;
- Design the data collection from administrative sources;
- Collect the data from the administrative and statistical sources;
- Implementation of data collection tools;
- Align the variables and measurements from the administrative sources;
- Integrate the data;
- Review and validate the administrative data;
- Edit and impute the data;
- Derive new variables and units;
- Calculate weights;
- Calculate aggregates and prepare the requested table;
- Use disclosure control;
- Document the processes used for creation of the statistical table.

Remarks on the workflow

Important issues which deserve special attention are:

Reconciliation of the information contained in the survey data and the administrative data.

Adaptation of the weights taking into consideration the mix of survey data and administrative data;

Specification of the imputation process, in particular consideration of mass imputation.

Usage scenarios for substitution and supplementation

Usage scenario: Split population approach

Input

A combination of administrative data (micro-data) and statistical survey data (micro-data). A part of the population is represented by the survey data and a part is represented by the administrative data. The two sub-populations are non-overlapping.

Output

Statistical products from the combination of administrative data and survey data.

Methodological options

Option 1: Pre-processing (Editing and imputation, Reviewing and validation) is done separately for the sources. Afterwards the two datasets are added and the weights in the survey data are adapted.

Option 2: First add the two datasets and apply the necessary sub-processes for the complete dataset.

Evaluation of the options

- Option 1 has some advantages from the work load but the disadvantage that inconsistencies cause by the complete dataset cannot be detected. Also with respect to the imputation it can happen that the complete dataset shows different correlations in the variables used for imputation.
- Option 2 is preferable because the different GSBPM sub-processes can use information from the complete dataset. The imputation is in this case closely related to the application of mass imputation.

For evaluation of the solution one can use the indicators for the different sub-processes described in the previous chapters of these Guidelines.

An additional indicator is checking whether there are some duplicates in the two datasets which violate the assumption of non-overlapping sources.

Usage scenario: Split variables approach

Input

Micro-data sources which cover the entire population, but variables in different datasets.

Output

Statistical products.

Methodological options

The options are in this case a little bit different to the previous case:

Option 1: Pre-processing (Editing and imputation, Reviewing and validation) is done separately for the sources. Afterwards the two datasets are integrated. Finally the complete dataset is processed according to the different GSBPM sub-processes.

Option 2: First add the two datasets and apply the necessary GSBPM sub-processes for the complete data set.

Option 3: Hybrid approach: Perform light pre-processing for the two data sets, afterwards integrate the two datasets and make a more detailed processing according to the different GSBPM sub-processes.

Evaluation of the options

- Option 3 seems to be the best compromise between the workload and quality of the output. In all cases, the light pre-processing should encompass all the variables which are used for matching of the two datasets.
- Option 1 is acceptable but means a lot of effort because of duplication of some activities.
- Option 2 should be avoided because the information used for matching is probably not of the highest quality.

The indicators for evaluation are defined according to the indicators for the sub-processes.

Examples

The split population approach is applied by many NSIs for the production of Structural business statistics. In this application the survey data are frequently a cut-off sample which does not consider enterprises with a small turnover. Using administrative information from tax registers allows information for all businesses to be included.

References

In its modules, the Memobust handbook of modern business statistics describes the different sub-processes in a lot of detail.

Manzi, Giancarlo, Spiegelhalter, David J., Turner, Rebecca M., Flowers, Julian and Thompson, Simon G., (2011), Modelling bias in combining small area prevalence estimates from multiple surveys, *Journal of the Royal Statistical Society Series A*, 174, issue 1, p. 31-50