Processing methods in business statistics (at national level)

Preface

This article describes the wide range of methods available to national statistical institutes for processing statistical inputs into statistical outputs.

The methods outlined here are mainly intended for the national statistical process, rather than for background data (e.g. data sources and the business register) or the further processing of compiled data to produce EU-harmonised statistics (e.g. standards on data validation, reference metadata reporting or dissemination).

It is also limited to methods which can be applied generally across statistical subjects. For **domain-specific methods**, please see the detailed domain methodologies.

You'll find a complete overview of methodologies and metadata for business statistics in the European Business Statistics manual.

Content

1. Introduction

The methods used to generate business statistics at national level reflect the various steps in the process: from the design of the process and determining and sampling the business population to data collection, error cleaning, processing missing values and calculating statistical output aggregates.

These steps typically follow the Generic statistical business process model drawn up by the United Nations Economic Commission for Europe (UNECE).

This article also examines methods for improving cross-domain consistency at microdata level between the various fields of business statistics.

Most of the methods summarised in this article are taken from the European Statistical System's (ESS) Handbook on Methodology of Modern Business Statistics, known as 'Memobust'. The sections below closely follow the various summaries in the handbook and provide links to several ESS methodological research practices.

2. Design of the process

The design of the statistical process generally refers to the design of a new survey, the redesign of a survey, or continuous improvements to a repeated survey.

The two main steps in the design process are:

1. **choosing methods**, e.g. sampling and estimation, data collection, contact strategies and editing

2. **allocating resources** to the subprocesses in compiling statistics.

In most cases, the design will be based on a particular statistical infrastructure (e.g. the business register, classifications, and types of data sources) and a particular set of statistical outputs.

It is essential to properly identify the variables of interest in the survey. These variables serve as an input for producing the statistical output and are not necessarily the same across Member States or over time, even though the statistical output as an end result is harmonised and consistent. This is part of the output-oriented approach of European business statistics. In this approach, it is up to the Member States to determine the best way of producing the European statistical output.

The aim of the design is to optimise statistical quality, minimise costs for the data compilers and the administrative burden on businesses, and maximise benefits for end users.

In practice, much of the design work is devoted to optimising the accuracy and reliability of the statistics at a more or less predetermined level of operational costs and under restrictive conditions regarding the burden on businesses.

There are additional important quality components such as timeliness, consistency and comparability. The optimisation process may include one or more of these components, often with certain trade-offs.

More information on design and optimisation can be found in the article on overall design (pdf). For specific guidance on the trade-off between accuracy and delays, see the guidelines for balance between accuracy and delays.

Most business statistics surveys are conducted at regular intervals – every month, quarter, year or over several years. These are referred to as 'repeated surveys'. The repetitive nature of a survey plays an important role in its design, as it affects sampling and accuracy, the perception of burden on businesses, time series and possible breaks. More information can be found in the article on repeated surveys (pdf).

In addition to the survey-oriented part, the design process also includes:

- A review of the existing and available administrative data – this can be very useful to check if a survey is really needed and for writing the questionnaire (by avoiding asking for information which is already available in administrative data).

  A detailed description of the various external data sources that can be used as input for the statistical production process, their usefulness and also their risks and recommended quality checks can be found in the section on 'Administrative data' in the article Data sources for business statistics.

- The domains of dissemination, including the level of detail, must also be identified in this step. The sampling design and the strategies for control and data validation all depend on the level of dissemination. The level of detail of the business statistics in the Framework Regulation Integrating Business Statistics (FRIBS) are described in Data requirements of business statistics.

- The units used to collect the input data and to disseminate the statistical output must be determined in this step. The units used for input can be different from the *statistical units* applicable to the output (e.g. legal units for input versus enterprises for output). The principle of subsidiarity in producing European business statistics enables Member States to use various types of input, provided that the resulting output from the statistical production process remains EU-harmonised.

- Data providers (or their representatives) should be involved early in the design process to assess the feasibility of the intended design (e.g. data availability) and also to create goodwill for the new design (e.g. involve business associations who might otherwise be unwilling to promote the survey).

3. Design of questionnaires

Questionnaire design is part of the operational phase of a survey, as it carried out after the questionnaire has been selected as the data collection method. However, it is critical in terms of the survey objectives.

It is difficult to compensate at a later stage for errors caused by an inadequate questionnaire (Brancato et al., 2006). As such, the design of questionnaires can be seen as essential to the design stage as a whole.

The relationship between information demand, response burden and existing information (the development of microdata linking) must be taken into account when creating new questionnaires or assessing existing ones. Questionnaire drafting, which is an iterative process, must be seen as a continuous cycle.

General information on designing questionnaires is available in Questionnaire design (pdf) and in the Handbook of recommended practices for questionnaire development and testing in the European statistical system (pdf).

There are also a number of more specific issues connected with drafting the questionnaire, including:

- designing electronic questionnaires (pdf)
- embedded functionalities for editing electronic questionnaires (pdf) that enable responding businesses to directly detect and correct any errors and/or missing values while filling in the questionnaire
- testing questionnaires (pdf).

As regards the embedded editing functionalities, receiving higher quality responses from businesses may significantly reduce the resources needed to clean the received microdata.

Testing the questionnaires is very important. Tests should be conducted at every stage in the process. It is good practice to have an advisory committee to take account of user needs and ensure that businesses will be able to answer the questionnaire.

In some countries, there are procedures for certifying survey quality (see, for example, Assessing and improving quality in official statistics: the case of the French Label Committee).

For more information on the testing and evaluating questionnaires see Brancato et al. 2006 and Willimack 2013.

4. Target business population (survey frame)

The survey frame identifies and lists the units of the business population together with their contact details, economic and geographic classifications and size classes. The sampling survey serves as a sampling frame.

The survey frame is also useful for contacting the data supplier and personalising and mailing the questionnaires. Furthermore, it has a role in the controlling and monitoring the data collection

phase, it helps to register and validate responses and evaluate non-response. The survey frame gives information for the weighting, grossing and micro-integration.

For business statistical surveys, the main source of the survey frame is the business register which records and maintains the statistical units and their characteristics. The business register can also store the links between units for collecting the data (i.e. reporting units) and units for dissemination (i.e. statistical units).

The survey frame for a particular survey 'instance' uses a snapshot of the register – the register state for a given date.

Since the business register serves as a base for different surveys, it is worth creating a master frame that can be used as a common frame for all surveys. A master frame and predefined subpopulations are useful for building survey frames and support the integration of different surveys.

Integrated survey frames improve the effectiveness of data collection and the whole survey process and also help to reduce response burden. As such, the *survey design* may pave the way for the integration of surveys by assigning suitable survey frames. It assigns the building blocks of the populations and the common classifications that might help to integrate data coming from different surveys.

Survey design can also accommodate the phenomenon whereby the information contained in the business register improves over time. This can be achieved by basing the collection on the initial register state, but later using the most recent state of the register (for the same reference period) for imputation and weighting.

A more detailed explanation of survey frames and their design can be found in Survey frames (pdf) and in Survey frame design (pdf).

5. Selecting samples

Sample selection in business statistics can be challenging for several reasons. The population is often skewed, new companies may be formed or go out of business, and businesses may be related to each other in different ways.

The use of *stratified simple random sampling* can enable researchers to draw inferences about specific subgroups that could be lost in a more generalised random sample, but it requires the selection of the relevant stratification variables.

A useful approach here, often used for business surveys where element sizes vary greatly, is to use probability proportional to size (pps) sampling, often combined with cut-off sampling.

This method can improve the accuracy of a given sample size by focusing the sample on large elements that have the greatest impact on population estimates. Stratification may also produce a more accurate estimate, especially if the strata regroup similar units. The cut-off method leads to bias, which must be quantified.

An alternative to stratified simple random sampling is *systematic sampling*. Cluster or multistage sampling is used for practical, economical and sometimes administrative efficiency. The use of fixed panels will produce very efficient estimates of periodic change. In most periodic surveys, sample rotation is used to reduce response burden.

A broad general introduction into these sampling techniques can be found in the articles on sample selection (pdf) and sampling issues in business statistics.

There are some cases where additional specific sampling techniques may be necessary.

For example if:

- the variable of interest is correlated to auxiliary variables that can be used in the design of the sample (see Balanced sampling (pdf)). This information can also be used with Neyman allocation based on the dispersion of the auxiliary variables.

- you need to produce preliminary estimates (see Subsampling for preliminary estimates (pdf));

If it is necessary to coordinate samples to produce comparable, consistent statistics, the estimates of change over time must be highly accurate and the response burden should be spread evenly between businesses (see Sample coordination (pdf)).

Coordination across different/sequential samples can be achieved by assigning permanent random numbers to the units in the business register.

There are two methods for sample coordination:

1. Simple random sampling with permanent random numbers (pdf)

2. Poisson sampling with permanent random numbers (pdf).

It is also possible to coordinate samples that are based on different statistical units (see Assigning random numbers when co-ordination of surveys based on different unit types is considered (pdf)).

If the units for collecting the data and for dissemination are different, you need to make some adaptations (see, for example, 'A first assessment of the impact of profiling on sampling', paper presented at Geneva ICES-V).

6. Data collection

The process of data collection involves a number of subprocesses, each with its own recommended methodology and specific considerations: the design phase of the data collection methodology, the techniques and tools for data collection and the mixed mode approach. This section focusses on methods relating to the following data sources:

- surveys
- reusing existing external data sources
- microdata linking.

6.1 Surveys

The choice of technique to depends on many factors, such as:

- survey subject
- timing of data delivery
- type of respondents
- budget.

The survey technique is usually chosen during the design phase, as the technique influences the way the data is collected and the design of the survey questionnaire.

There are various [techniques and tools for data collection](#) (pdf).

For example:

- computer-assisted telephone interviewing (CATI)
- computer-assisted personal interviewing (CAPI)
- e-mail and online surveys
- the electronic exchange of information based on electronic data interchange (EDI) and eXtensible business reporting language (XBRL).

By uploading data files in a standard record layout, perhaps integrated into a web questionnaire, you can obtain high quality data with a relatively low response burden.

The use of the mixed mode approach, i.e. combining different data collection techniques in the same survey, can overcome the limitations specific to each technique. If the approach is designed correctly, it can reduce the unit non-response rate.

The data collection process concerns not only interviewing techniques, but also contact strategies, monitoring activities and follow-up:

1. **Contact strategies** are necessary to get in touch with respondents and may vary according to the type of respondent unit (large or small company, new company, etc.).

2. **Monitoring activities** are important to keep the data collection process under control while it is in progress and to take proper action to improve or modify any factors that could seriously impair data quality.

3. **Follow-up** takes place after the formal data collection period has ended. It involves following up on non-respondent units and the strategy for doing so (based on their significance on statistical end results).

6.2 Reusing existing external data

A general trend among the national statistical institutes is to reuse administrative data already collected by other public organisations or other existing external data sources, including big data. It also includes data sources from other statistical institutes, as in the case of [microdata exchange for Intrastat](#).

These external data sources are generally referred to as 'secondary data' as opposed to 'self-collected' data (i.e. 'primary data').

The most obvious advantage of reusing existing information is a reduction in (collection) costs and the burden on business. However, there are various pros and cons to be taken into account when deciding on the methods for collecting and using secondary data (pdf).

6.3 Microdata linking (data fusion)

For some statistical elements, you can avoid collecting primary or secondary data by combining existing (internal and external) microdata sources.

This approach is known as 'microdata linking' or 'data fusion' and involves various techniques for integrating several, sometimes conflicting, microdata records into a new set of high-quality microdata records.

In addition to the general overview of data fusion at micro level (pdf), you'll find more detailed technical information below, depending on the quality of and overlap between the microdata sources:

- If the statistical units (or other record identifiers) represented by the different microdata sources more or less match, use object matching (pdf). If the match is of good quality, see object identifier matching (pdf). For matches of poor quality, consider alternative methods, such as unweighted matching of object characteristics (pdf), weighted matching of object characteristics (pdf), probabilistic record linkage (pdf) and the Fellegi-Sunter and Jaro Approach (pdf).

- If there is no overlap of record identifiers between microdata sources — e.g. when using different statistical units — even though the sources target the same population, the recommended methodologies are more complex and are referred to as statistical matching methods (pdf). See also general background information on statistical matching (pdf).

Once the new microdata set has been created using one of these matching techniques, the new set may contain conflicting microdata. You'll find a general description of this problem and how to resolve it in the article on reconciling conflicting microdata (pdf). For more specific reconciliation techniques, see point 3 of section 7.

Finally, if the data collection units differ from the statistical units, the last step is to consolidate the answers of reporting units.

7. Checking and cleansing microdata

After collecting the microdata using surveys, existing external data sources or microdata linking of existing internal sources (see section 6), you need to check and clean the microdata records.

This process is referred to as 'editing' (for a detailed overview see Statistical data editing (pdf) and Recommended practices for editing and imputation in cross-sectional business surveys (pdf)).

The checking and cleansing methods include several techniques that can be used together or separately:

1. deductive editing (pdf): for treating systematic (recurring) errors throughout the dataset;

2. selective editing (pdf): mainly for treating specific micro-records, e.g. those of larger enterprises;

3. automatic editing (pdf): for treating errors that can be fully edited automatically. Special editing techniques are available in case of conflicting microdata that has been 'collected' by means of microdata linking (see chapter 6.3), such as prorating (pdf), minimum adjustment (pdf) and generalised ratio adjustments (pdf);

4. manual editing (pdf): for treating errors using expert judgment. Because of its relatively labour-intensive nature, it is often accompanied by well-defined editing instructions and restricted to those errors which have significant impact on the outcome and could not be treated by other editing techniques;

5. macro-editing (pdf): for treating only those errors that would have a significant impact on the (aggregated) statistical output data.

    This technique is therefore also known as 'output editing'. To ensure consistency between the microdata and the final aggregated outcomes, these significant errors are corrected at microdata level and not at statistical output level. This type of editing in the field of business statistics differs from the *macro-integration* used in national accounts to balance supply and use at national level.

In the case of the use of secondary information in general and administrative data in particular, some additional specific editing considerations may apply, see Editing administrative data (pdf).

Additional editing techniques for *time series* are described in Editing for longitudinal data (pdf).

During the checking and cleansing of the microdata, data compilers at national level can already apply the data validation standards for output data transmitted to Eurostat.

Incorporating these standards into regular data checking and cleansing routines reduces the risk of data being rejected by Eurostat. Although the EU data validation standards can be incorporated into the cleansing of microdata, they are different, as microdata cleansing focuses on source data from the Member States, whereas EU data validation focuses on data transmitted by the Member States to Eurostat.

8. Imputing missing values

The problem of missing values occurs both for data collected in traditional surveys and for administrative data.

It is usually more difficult to use an incomplete dataset to infer population parameters, such as totals or means of target variables. For this reason, data compilers often create a complete dataset prior to

the estimation stage by replacing the missing values with estimated values from the available data. This process is referred to as 'imputation'.

Possible imputation methods include:

1. deductive imputation (pdf): this method is used if the missing value can be logically calculated from available non-missing values, e.g. in the case of a missing total and its non-missing sub-totals;

2. model-based imputation (pdf): this method is based on a predictive model based on the quantitative relationship between the missing value and observed non-missing values;

3. donor imputation (pdf): the missing value is imputed by a 'donor' record with non-missing values and with similar characteristics.

In order to comply with the editing rules it is necessary to constrain the imputation methods either directly (complex) or stepwise (simpler). For more details see Imputation under edit constraints (pdf).

Different methods may be appropriate in different contexts.

Some general aspects of imputation that do not relate to a particular method, such as the inclusion or exclusion of an error term in the imputed values, the use of deterministic versus stochastic imputation, the incorporation of design weights into imputation methods, and multiple imputation and mass imputation, are discussed in Imputation – main module (pdf).

There are alternative methods for dealing with non-response in addition to those described above, although they are more complex. These alternative methods rely on reweighting procedures that are integrated into the methods for estimating aggregated totals (see section 9).

9. Estimating aggregated totals (output data)

After the microdata has been cleansed (see section 7) and imputed for non-response (see section 8), the next step in the compilation process is to estimate aggregated totals from the observed microdata.

This section gives an overview of the methods that can be used to obtain estimates for parameters such as aggregated totals, means and ratios.

A general overview of estimation methods and how to design estimation can be found in Weighting and estimation (pdf) and in Design of estimation (pdf).

The estimation methods can be divided into *design-based* (traditional) and *model-based* approaches. Model-based estimation methods are used if there is no random sample design available underpinning the microdata (e.g. in the case of data from incomplete administrative sources or from an unknown internet source) or if there are too few observations to produce reliable estimates by means of the traditional design-based estimators.

Commonly, in official statistics, probability-based sampling designs are used, and a design weight can be associated with each sampled unit. This design weight equals the inverse of the inclusion probability. It can be thought as the number of population units each sample unit represents.

Therefore, a simple method for obtaining estimates of the target parameters is to use these design weights to inflate the sample observations. Design weights are strictly linked to the sampling design used for the survey. Moreover, design weights can be adjusted to consider non-response or they can be modified to take account of auxiliary information.

An example of use of external information is provided by the estimator based on calibration (pdf) or on generalised regression (pdf) which is a special case of a calibration estimator.

In the case of non-response, several methods are available — based on adjusting design weights — that take into account (temporary) non-response as an alternative to micro-imputation of missing units as described in chapter 8. For general methods that can be used if the theoretical sample is not achieved in the observed sample due to non-response, see Preliminary estimates with design-based methods (pdf).

The previous estimators are unbiased or approximately unbiased in a randomisation approach (in a design-based approach, the properties are assessed against the set of all possible samples).

Note that even if, in some cases, a model is assumed (as for generalised regression), the properties of the estimators do not depend on the model and the estimators remain design-unbiased even in the event of model failure. For this reason, this class of methods is robust. However, their efficiency depends heavily on model assumptions and relationships, and auxiliary variables can affect their variances.

In fact, if the distribution of the target variable in the population is highly skewed, as often happens in business surveys, representative outliers may appear in the sample. The values of such units are true values, so they do not need to be edited. Nevertheless, even if estimators remain unbiased, the presence of these outlying units has a major impact on variance estimators. See Outlier treatment (pdf) for an overview of methods that have been suggested for reducing the variance of the estimates while controlling for the presence of bias.

Model-based estimators can be applied in specific situations where the traditional design-based methods fall short.

This could be the case, for example, if the sample size is not large enough to obtain sufficiently accurate estimates. For general information see Small area estimation (pdf). More detailed technical information on the various small area estimations methods can be found in: Synthetic estimators (pdf), Composite estimators (pdf), EBLUP area level estimators (pdf), EBLUP unit level estimators (pdf), and Time series data estimators (pdf).

Methods specifically relating to administrative data can be found in Estimation with administrative data (pdf).

If the confidentiality of the aggregated totals is an issue, please see Statistical disclosure control.

10. Improving cross-domain comparability and consistency

In the design and compilation phase there are a number of ways to improve comparability and consistency across different statistics.

The coordinated use of the business register as the source to define the population and design coordinated samples is the first step in establishing comparable statistics.

A balance should be struck between sample size (and the associated administrative burden and compilation costs) and the expected accuracy of the resulting output data and its estimation method.

Maximising the reuse of previously collected data would also increase comparability. A key example is the reuse of VAT records for both annual and short-term turnover statistics.

The data compiler can also introduce a number of cross-domain checks at the micro-level for larger enterprises or enterprise groups, ensuring consistent microdata for those units that usually have a large impact on the final output data. This is usually referred to as 'micro-integration'. For an example of micro-integration at Statistics Netherlands, see here.

At the end of the compilation process, it is strongly recommend to add a validation step in which the resulting output data is confronted with comparable output data from other sources. This will enable data compilers to check that the strategy for control, correction and imputation was effective (see also data validation, especially validation level 4 which refers to cross-domain checks). This type of validation may also involve output checks that would support the integration process of national accounts for which the business statistics serve as an input.

At the French National Institute of Statistics and Economic Studies (INSEE), for example, the integration of structural business statistics (SBS) into national accounts is supported by a special validation procedure:

1. First, the previous SBS/NA-aggregates from year N-1 are recalculated using the SBS/NA aggregation method for year N. The method at year N may integrate improvements (consequently changes) compared with the method used previously in year N-1. The recalculated N-1 SBS/NA aggregates (based on the method for year N) may therefore differ from the original N-1 aggregates (based on the method for year N-1).

2. Second, the evolution of the recalculated SBS/NA aggregates using N-1 method and the SBS/NA aggregates for N (both based on the same method for N) are checked and validated for national accounts purposes.

11. See also

- Overview of methodologies of European business statistics: EBS manual
- Legal aspects related to this chapter can be found in the following overview.
- data sources
- microdata exchange for Intrastat
- business register
- data validation
- reference metadata reporting

- [Data requirements of business statistics](#)
- [principle of subsidiarity in producing European business statistics](#)
- [dissemination](#)
- [Statistical disclosure control](#)
- [detailed domain specific methodologies](#)

12. Further Eurostat information

- [Handbook on Methodology of Modern Business Statistics](#)
- [ESS methodological research practices](#)
- http://ec.europa.eu/eurostat/ramon/statmanuals/files/Handbook_of_Practices_for_Quest.pdf

Design of the process

- [overall design](#)
- [guidelines for balance between accuracy and delays](#)
- [repeated surveys](#)

Design of questionnaires

- [Questionnaire design](#)
- [Handbook of recommended practices for questionnaire development and testing in the European statistical system](#)
- [designing electronic questionnaires](#)
- [editing electronic questionnaires](#)
- [testing questionnaires](#)
- [Assessing and improving quality in official statistics: the case of the French Label Committee](#))
- [Brancato et al. 2006 and Willimack 2013](#)

Target business population

- [Survey frames](#)
- [Survey frame design](#)

Selecting samples

- [sample selection](#)
- [sampling issues in business statistics](#)
- [Balanced sampling](#)
- [Subsampling for preliminary estimates](#)
- [Simple random sampling with permanent random numbers](#)
- [Poisson sampling with permanent random numbers](#)
- [Assigning random numbers when co-ordination of surveys based on different unit types is considered](#)
- [A first assessment of the impact of profiling on sampling](#)

Data collection

- techniques and tools for data collection
- collecting and using secondary data
- data fusion at micro level
- object matching
- unweighted matching of object characteristics
- weighted matching of object characteristics
- probabilistic record linkage
- Fellegi-Sunter and Jaro Approach
- statistical matching methods
- statistical matching
- reconciling conflicting microdata

Checking and cleansing microdata

- Statistical data editing
- Recommended practices for editing and imputation in cross-sectional business surveys
- deductive editing selective editing
- automatic editing
- prorating
- minimum adjustment
- generalised ratio adjustments
- manual editing
- macro-editing
- supply and use
- Editing administrative data
- Editing for longitudinal data
- data validation standards

Imputing missing values

- deductive imputation
- model-based imputation
- donor imputation
- Imputation under edit constraints
- Imputation – main module

Estimating aggregated totals

- Weighting and estimation
- Design of estimation
- calibration
- generalised regression
- Preliminary estimates with design-based methods
- Outlier treatment

- [Small area estimation](#)
- [Synthetic estimators](#)
- [Composite estimators](#)
- [EBLUP area level estimators](#)
- [EBLUP unit level estimators](#)
- [Time series data estimators](#)
- [Estimation with administrative data](#)

Improving cross-domain comparability and consistency

- [The treatment of large enterprise groups within Statistics Netherlands](#)

13. External links

[Generic statistical business process model](#)

14. Contacts

For questions or comments on this chapter, please contact ESTAT-EBS-MANUAL@ec.europa.eu.