



This module is part of the

# Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

# Theme: Quality of Statistical Registers and Frames

## Contents

- General section ..... 3
  - 1. Summary ..... 3
  - 2. General description..... 3
    - 2.1 Importance..... 3
    - 2.2 Quality of statistical registers ..... 3
    - 2.3 Quality of the master frame ..... 8
    - 2.4 Quality of frames ..... 9
    - 2.5 Interaction between quality of statistical registers and survey frames ..... 11
  - 3. Design issues ..... 12
  - 4. Available software tools ..... 12
  - 5. Decision tree of methods ..... 12
  - 6. Glossary..... 12
  - 7. References ..... 12
- Interconnections with other modules..... 13
- Administrative section..... 14

## **General section**

### **1. Summary**

The quality of the statistical registers and survey frames is a fundamental part of the quality of the surveys built on these registers and frames. The quality can be assessed by the general European output quality components like relevance, timeliness and punctuality, accuracy, accessibility and clarity, coherence and comparability. Even though these components are typically used to measure the quality of statistical outputs, the quality of registers and survey frames can also be assessed using these components. For this purpose, the interpretation of these components in this context is also given in this module. Some of these criteria have more effect on the quality of registers and frames than others; these are timeliness and accuracy (the completeness, coverage of the population, precise and valid contact information and classification and size attributes of the units). Coherence and comparability of registers and survey frames are also significant to substantiate the integration of different surveys and administrative sources.

### **2. General description**

#### *2.1 Importance*

This module aims to serve as a link between the quality criteria, see the general topic on “Quality Aspects”, and the roles of statistical registers and frames, as described in the previous modules of the topic “Statistical Registers and Frames”. Quality has a prominent role in the statistical data collection as the whole production process is designed in a way to assure the collection of data with high quality. As it is described in section 2.3, statistical registers can be used as inputs for this process as bases of frames. They can also be the final products of the process and serve as bases for analyses.

From the point of view of data collection, the quality of the statistical register is of the utmost importance as it is used as the main source for the whole data collection process; thus it basically determines the quality of the data collection. The outcome of this process should also be quality-proof as feedback with all the errors of the survey frame and relevant paradata of the data collection should be passed on to the register and the erroneous content of the register should therefore be corrected.

The quality of statistical registers and frames (master frames and survey frames) can be investigated separately, even though there are strong links between them as the relationship between registers and frames is also close (see section 2.3 and 2.4). This interaction between registers and frames is often reflected in key points of assessing their quality as the quality of registers usually determines the quality of frames and vice versa. The next three sections will identify the key points in quality of statistical registers, master frames and frames separately and the last section briefly summarises the link between them.

#### *2.2 Quality of statistical registers*

Talking about quality usually comes with mentioning quality criteria. These criteria are used to characterise quality aspects in each phase of the statistical production process. These criteria – while still keeping their universal meaning – can be further clarified to apply to issues related to statistical registers and frames.

### 2.2.1 *Relevance*

**Relevance** can be interpreted as the adequacy of the register (as a statistical output) to user needs. These users can typically be divided into two groups: internal and external users.

Needs of *internal users* are usually expressed towards the register in ways that the register should serve the needs of the relevant data collections, e.g., provide all necessary information for the survey frames. The adequacy of the statistical register to internal user needs generally depends on whether it provides complete information to build survey frames, to access, stratify and group data suppliers, reporting units, to support sampling by providing all the variables at the needed level of detail. To fulfil *external needs*, the register should be harmonised with other registers, in order to move towards a common statistical register for the whole governmental administration.

Relevance of registers against their needs should always be taken into account to improve quality but various needs might be weighted differently considering their importance.

Registers are compiled to serve the needs of data collection. Therefore the relevance of a given register should be evaluated considering its adequacy to meet the expectations expressed towards it by the data collection system. The needs of main users should be explicitly expressed (written down) and the register should be regularly assessed whether it fully fulfil the justified expectations. The needs of external users can be identified and assessed using regular user opinion surveys. Lessons learnt from these surveys should always come back to register administrators and be implemented if possible to further improve their quality.

For the assessment of relevance of a register, the following information should always be collected and considered:

- exact definitions of register populations and units
- identification and clarification of main internal and external users of the register
- clarification of user needs: the attributes they are interested in (both collected and derived information) and the percentage of their needs that can be met using register data
- assessment of internal and external user satisfaction and the main lessons learnt from them

### 2.2.2 *Timeliness*

**Timeliness** of register information is considered as the time gap between the actual and registered state of register information. Register information is not always updated immediately after a change in units or their characteristics but only after a defined period of time. This periodicity should be minimised as it carries certain level of uncertainty whether the information used is the current data.

Timeliness of register shall be checked, considering the following aspects:

- the conformity between the reference period of data collections built on registers and the periodicity of register maintenance. In order to have more up-to-date information available from data collections, the maintenance periodicity of the register should not be less frequent than the periodicity of collection of new information

- the time interval between the changes  $c$  in the population occur (appearance of new units, death/cessation of units, changes in unit attributes, etc.) and their reflection in register information. This time lag should be minimised as shorter lag means data with less uncertainty
- the average number or percentage of new and ceased units during the time gap. The number of changes has influence on the demanded frequency of maintenance. Frequent changes demand frequent maintenance to assure the timely state of the register.
- the connection between the frequency of register maintenance and the reference times demanded by survey frames, statistical processing and analysis.

### 2.2.3 Accuracy

**Accuracy** is assessed as the level of completeness and compliance of register information with the actual real state of units of characteristics. There might be several causes of differences in accuracy (EUROSTAT, 2010):

- gaps in declaration procedures, certain events are not declared by the enterprises, they are not prescribed for the enterprise or they ignore their obligation. For example:
  - the change in their addresses or
  - the cessation because the entrepreneur does not know whether the cessation is final or temporary
- inaccurate declarations, these may be deliberate or unintentional. For example:
  - at reporting to the tax office the incomes may be under-recorded, and/or expenses over-recorded or
  - the inaccurate declaration of the principal activity because of the misinterpretation of the concept, administrative advantages/disadvantages, unknown future activity at registration, etc.
- coding errors typically at the classification attributes and size categories, that is in the stratification attributes that causes problems at sampling and estimations
- falsely active units that ceased trading. Being selected into the sample, they increase the cost and nonresponse

These differences are considered as errors and provide information on the accuracy of the given information. The most common sources of accuracy problems are coverage issues, classification errors and contact problems. To measure these accuracy problems, different methods can be applied:

- one way to assess the accuracy of the register is to consider the survey frames built on the register. As each of these frames usually covers a part of the register, the quality of the register can be inferred by considering all the feedback collected from the different surveys built on frames of the register
- control surveys to measure register accuracy in contact information, classification by activity or size and an estimation of the proportion of falsely active units

- compare the register information with external sources, information from other register or other administrative data. For example:
  - Comparison of the number of employees from the business register and the social security register can indicate coverage problems
  - In order to control contact information, telephone registers can be used
- other methods, e.g., benchmarking, process quality audit, etc.

The accuracy of the coverage can be measured against the target population of the register. "In principle, business registers should record all enterprises (and associated legal and local units) and all truncated, multinational and all-resident enterprise groups that are active in the national economy, i.e., contributing to the gross domestic product (GDP)" (EUROSTAT, 2010, chapter 6). In practice, certain groups can be omitted (like households, very small enterprises, etc.) because of costs and international comparison or other reasons. The business register regulation (EC, 2008) describes the scope and unit types that are compulsory or recommended to be in the national business register. The coverage has to be measured against the target population of the given register, not against the population of interest (the ideal target population).

The coverage error might contain units outside the scope of target population (over coverage) or omit units that belong to the target population (under coverage) (see the module "Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys").

- over coverage is caused by multiple registrations of units or the missing declaration of the changes of the activity status (dead, ceased, inactive units).
- under coverage is a result of missing or delayed registration of units

Coverage issues in registers should be evaluated considering the following aspects:

- information on any difference between the target population and frame population
- if there are coverage issues, the causes of these problems should be collected and organised to identify the real causes

*Misclassifications* occur when information on classification, size category characteristics of units is missing or false. In business register, errors in classifying the statistical main activity (NACE), geographical position, legal form, staff and turnover categories are typically affected by this kind of accuracy problem. Classification concerns should be analysed by thinking of:

- identifying the characteristics or subpopulations of registers usually affected by this kind of error
- quantifying the misclassifications by counting their number or ratio
- measuring the number of revisions and changes made (not considering scheduled register maintenance)

*Contact problems* usually result in unsuccessful contacts with the data provider. Contact information is the name of the unit, the seat and postal addresses, phone, fax numbers and e-mail address, the name, position and accessibility of the contact persons and responsible managers, etc.

Contact problems should be registered to give a base for the measurement, and for actions to find ways to improve unit accessibility. Information for the analysis of contact errors can be gathered by

- checking the postal logs and the error codes received
- collecting information on nonresponse of the surveys based on the register

#### 2.2.4 *Accessibility and clarity*

**Accessibility and clarity** criteria usually refer to the circumstances under which internal and external users access information. In case of registers, the ways that users get access to the data stored in the register are considered. Access conditions can be further analysed by the range and level of detail of register attributes the users have access to. This criterion can be assessed by:

- identifying the ways external users access register information. Modes and conditions of access should also be evaluated
- mapping the channels that internal users use to access register information. Modes and conditions of access should be evaluated for internal users as well
- checking the availability and detail of register documentation. Available documentation for external users can highly improve accessibility and clarity of the registers. Update intervals of register documentation should be in line with changes in register methodology

#### 2.2.5 *Comparability and coherence*

**Comparability and coherence** can be interpreted both within the scope of the given register or between registers. Analysing coherence between registers shows whether registers with similar content can be linked using a common identifier or key variables. Coherence within the scope of the given register means unified change management and logical connections between characteristics and/or units. Comparability in time refers to the relationship between information on the same register units and characteristics in different periods of time. The geographical comparability usually has meaning at higher geographical (typically international) levels. The same interpretation of the concept of units and classifications in different geographical spaces can improve comparability of information derived from the registers. Comparability with external register sources can also serve the needs of effective update procedures of the statistical register. This criterion of comparability and coherence can be analysed by thinking of:

- the level of compliance with standards and recommendations of the European Union or the United Nations or other international organisations;
- the length of time series of data and analysing the latest breaks in the time series and their causes (e.g., changes in classifications, methodology, etc.);
- the possible logical connections between register variables;
- the usability of register sources in forms of common terms, units and maintenance principles. Standard classifications and nomenclatures should be used in different registers to promote comparability. In relations with other registers connected to the statistical register, consistency should be ensured at the level of statistical units.

Statistical registers shall also be suitable to connect to administrative registers. This connectivity creates the basis to integrate administrative information into the statistical production system. Without harmonising statistical registers with administrative sources, administrative information cannot be an integral part of the process.

### 2.2.6 *Quality of statistical registers - Summary*

Considering the key points of the previous modules of the topic “Statistical Registers and Frames” and the quality aspects described previously, the following summarising principles and quality directives can be defined for the quality aspects of statistical registers:

- registers should be built and maintained to coordinate data collections with common target populations
- standard classifications and nomenclatures shall be used in different registers not just to support comparability but also to achieve effective maintenance procedures
- contact information should be cross-checked between registers (postal codes, address register)
- the date of effect and change information should be used to select the frame (master frame) units and their attributes to a given reference time
- interoperability between different registers shall be assured to the highest possible extent when establishing registers
- standard procedures should be available for change management (appearance of new units, death/cessation of units, changes in unit attributes, etc.). Keeping track of the reasons, scope and sources of changes and the traceability of historic data should also be ensured
- as many available administrative data as possible shall be used for register maintenance. Feedback from statistical data collections should also be made
- registers shall be updated as frequently as possible with special attention given to the needs of users and survey frames
- coverage errors shall be minimised, i.e., units stored in the register should correspond to the target population to the highest possible level and units outside the scope of the target population should be omitted from the register
- special attention should be given to the key units in forms of regular verification, control and profiling to improve the quality.

### 2.3 *Quality of the master frame*

The master frame, or in other words, survey universe file is a snapshot of the business register (and other supplementary or satellite registers). Its quality aspects are similar to the registers used as base for the master frame. For the following aspects, the interpretation of quality of the master frame is different from register quality:

- the relevance from the point of view of the usage of the master frame for different surveys
  - The quality of the master frame gets better with more types of survey frames selected from it

- The scope of the frame population can be defined by the conditions on the frame attributes and/or by other lists of the units. The quality of the master frame is better if the level of information in the master frame is adequate to its assignment to the highest possible extent
- the coherence and comparability of the survey frames built on the master frame. The design of frames and samples has to prepare the integration of data from different surveys. The possibility and result of record linkage is a peculiarity of the quality of master frame and the survey frame design (see the module “Statistical Registers and Frames – The Design of Statistical Registers and Survey Frames”)

#### 2.4 *Quality of frames*

Similarly to the quality of registers, the quality criteria can also be further specified for survey frames.

**Relevance** from the survey frame point of view can be considered as the level of adequacy to the needs of a given survey using a given survey frame. In order to meet their expectations, the needs of different surveys should always be precisely expressed.

To assess the relevance of a survey frame, one should think of:

- the availability of different types of units of the survey (reporting, collection, etc. units)
- definitions, classifications, contact information required by the survey based on the given frame
- the regular assessment of users of the frame and frequent feedback about their needs to frame designers

**Timeliness** of the frame is closely linked to the register the frame is built upon. Timeliness of a frame means that the frame is in line with the reference time of the data collection. Any difference between them indicates potential timeliness issues, meaning that the time gap between the actual state of units, their characteristics and their given state in the frame indicates such problems. In order to minimise this time gap, the survey frame must be drawn from the snapshot of the register that is suited to the reference time of the survey. One possible solution is the Hungarian one, where for annual surveys, the snapshot of the register of December 31<sup>st</sup> is proposed to be used. For monthly surveys, twelve survey frames have to be created in a year for each survey instance; each has to use the monthly snapshot of the register that is latest to the data collection period.

Timeliness of the survey frame can be evaluated against the following criteria:

- the gap between the reference time of the frame and the survey. The time lag will introduce bias with timeliness issues
- the period of time that a change in population (appearance of new units, death/cessation of units, changes in unit attributes, etc.) is reflected in the frame information. This criterion highly builds upon the quality of the register the frame is based on

**Accuracy** is considered as the level of completeness and compliance with the frame information with the actual state of units or their characteristics. Just like for registers, there might be several causes of differences in accuracy of frames: coverage issues, misclassifications and contact problems.

*Coverage issues* mean that survey frames contain units outside the scope of target population (over coverage) or missing units that belong to the target population (under coverage). Over coverage issues of the frame can be analysed considering the feedback from survey control (number of units that should have been excluded from the survey) or from consultations with data providers. In order to analyse over coverage at the sufficient level, adequate nonresponse codes should be used in survey control processes. Examining under coverage is more difficult as acquiring information from indirect sources or questionnaires received from non-surveyed units is generally the basic method to measure under coverage. The causes of coverage errors, additional to the register coverage errors, can be:

- the misclassification of register units in the attributes referred in the conditions of assignment of the scope of the frame population
- the external lists of units used to select the frame population (e.g., units with the observed activity type)
- the timeliness error mentioned above
- the level of detail required by surveys but not covered by frames. This type of problem might occur if the frame has missing information at the required level of detail thus only information at higher aggregated levels can be collected (e.g., if information at local unit level is sought but it can be provided only at enterprise level due to lack of information).

Frame coverage issues should be evaluated considering:

- the information on any difference between the target population and frame population
- the coverage issues of the register the frame is built upon. In case the variables that are used for the creation of the frame are erroneous in the register, coverage issues are likely to occur for the frame as well

*Misclassifications and contact errors* are special types of biases for frames in the sense that they highly depend on the input registers. These types of errors can be measured by the quality of the given register subpopulation.

One part of the accuracy problems can be corrected during the surveys if:

- missing units can be added to the frame, given that they emerge and exist in the base register (business register)
- redundant or unnecessary units can be omitted after the verification of over coverage
- the classification or the address errors can be corrected with feedback given to the statistical registers

The improvement of the survey frame is supported by a suitable classification of nonresponse codes. Nonresponses have to be declared during the data collection phase and can be considered for the next instances of the survey. If common nonresponse codes are applied for the master frame, then not only the given survey, but the frame of other surveys with common frame units will become more accurate.

**Accessibility and clarity** criteria can be interpreted as the frames are designed in a way that they are used for the purposes of survey design. Accessibility criterion is considered that there are no obstacles to access the frame. The content and way of creation of the survey frame from the register (master frame) should be documented.

**Comparability and coherence** in frames means that logical connection between characteristics and/or units are checked and assured. It can be interpreted either within a given survey frame or between other frames. Within the survey frame, it takes over the quality problems of the base register. Among the survey frames, the common identification, subpopulations and classifications substantiate the integration of the different surveys. In order to assure a higher level of comparability and coherence between frames, the populations of the data collections should be defined so that they support both horizontal and vertical comparisons.

### **Quality of frames - Summary**

Considering key points of the previous modules of the topic “Statistical Registers and Frames” and the quality aspects described previously, the following summarising quality directives can be defined for the quality aspects of frames:

- survey frames shall correspond to the target population to the highest possible extent. Over and under coverage should be minimised
- the quality of the potential registers used for the design of data collection shall always be taken into account and evaluated against quality criteria
- survey frames shall always contain the latest information available according to the reference time of the survey
- identifiers, contact, classification and size attributes of units shall always be up-to-date as stratification, selection of samples, data processing, imputation, estimation, data linkage and matching, quality evaluation and analyses are based on this information
- special attention should be given to frame errors (coverage issues and outdated information) as they might distort survey results, deteriorate quality and increase costs
- in order to link data collections and survey frames they are built on, common subpopulations for these data collections should be defined
- coverage of the survey frame shall regularly be assessed, therefore adequate nonresponse codes should be defined for the data collection process to derive information for coverage issues at adequate detail
- supplementary information is to be sought and used to manage differences between survey frames and target populations
- the documentation of survey frames shall always contain the description of target population, survey frame and coverage of the given survey

#### *2.5 Interaction between quality of statistical registers and survey frames*

As seen in previous sections the relationship between quality of registers and frames could be really close. Accuracy and timeliness are typical quality criteria where distinction between the two cannot easily be made. Clear causal relationship between them sometimes cannot be drawn as errors identified at survey frame level might be closely related to the similar issues with the registers the frame is built on.

The quality of the survey frame is in turn an important component of the quality of the whole survey. Therefore the quality aspects of registers, frames and data collections should be considered as a whole. The improvement of the register quality contributes to the improvement of the quality of the survey frame, the data collection, and the output of the survey. From the errors of the frame feedback can be derived to the register, so it can improve the quality of the register as well.

### **3. Design issues**

### **4. Available software tools**

### **5. Decision tree of methods**

### **6. Glossary**

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

### **7. References**

European Parliament and the Council of the European Union (2008), Regulation (EC) No 177/2008 of the European Parliament and of the Council of 20 February 2008 establishing a common framework for business registers for statistical purposes and repealing Council Regulation (EEC) No 2186/93. *Official Journal of the European Union*, L 61, 05/03/2008,0006–0016.

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:061:0006:01:EN:HTML>

Eurostat (2009), *ESS Handbook for Quality Reports, 2009 edition*.Eurostat Methodologies and Working papers. [http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-EHQR\\_FINAL.pdf](http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-EHQR_FINAL.pdf)

Eurostat (2010), *Business Registers - Recommendations Manual, 2010 edition*.Eurostat Methodologies and Working papers. [http://epp.eurostat.ec.europa.eu/cache/ITY\\_OFFPUB/KS-32-10-216/EN/KS-32-10-216-EN.PDF](http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-10-216/EN/KS-32-10-216-EN.PDF)

Statistics Canada (2003), *Statistics Canada Quality Guidelines - Coverage and frames*, 17–20.<http://www.statcan.gc.ca/pub/12-539-x/index-eng.htm>

## **Interconnections with other modules**

### **8. Related themes described in other modules**

1. General Observations – Methods and Quality
2. Statistical Registers and Frames – Main Module
3. Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys
4. Statistical Registers and Frames – Building and Maintaining Statistical Registers to Support Business Surveys
5. Statistical Registers and Frames – Survey Frames for Business Surveys
6. Statistical Registers and Frames – The Design of Statistical Registers and Survey Frames
7. Statistical Registers and Frames – The Statistical Units and the Business Register
8. Quality Aspects – Quality of Statistics

### **9. Methods explicitly referred to in this module**

- 1.

### **10. Mathematical techniques explicitly referred to in this module**

- 1.

### **11. GSBPM phases explicitly referred to in this module**

1. Quality management – overarching

### **12. Tools explicitly referred to in this module**

- 1.

### **13. Process steps explicitly referred to in this module**

- 1.

## Administrative section

### 14. Module code

Statistical Registers and Frames-T-Quality

### 15. Version history

Version	Date	Description of changes	Author	Institute
0.2	29-05-2012	first version as a separate module	Zoltán Vereczkei	HCSO
0.3	21-06-2013	change of template	Ildikó Györki	HCSO
0.4	17-11-2013	revised version after EB review	Zoltán Vereczkei	HCSO
0.4.1	11-12-2013	preliminary release		
0.5	30-01-2014	revised version after EB second review	Zoltán Vereczkei	HCSO
1.0	26-03-2014	final version within the Memobust project		

### 16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:38