This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Data Collection – Main Module

Contents

General section

1. Summary

Data collection is a "*systematic process of gathering data for official statistics*" (SDMX, 2009).

It is a very articulated process that develops itself along different steps of the survey process: from the design phase of the data collection methodology through the finalisation of the collected information (GSBPM, 2009), in order to collect data for statistical purposes by using many different techniques that can or cannot be assisted by computer and can or cannot need the support of interviewers (main ones: CAPI, CATI, WEB, PAPI, mail questionnaires and direct observation).

The choice of the technique to use depends on many factors (survey theme, timing of data delivery, difficulty in founding the information required, type of respondents involved, budget, etc.) and it is generally taken during the design phase of the process since the technique influences the way the data collection is carried out as well as the design of the survey questionnaire.

The use of mixed-mode, that is the combination of different data collection techniques for the same survey, can overcome those limitations that are specific of each technique and, if correctly designed, can reduce the unit non response rate.

A general trend among the NSIs is to gather the information they need by using administrative data in order to reduce respondent burden as well as costs. This is because NSIs can take the advantage of using already existing data, stored in public archives hold by other public organisations that have already performed a "data collection" phase, according to their needs and purposes that, anyway, might differ from the statistical ones. This trend is helped by the IT rapid developments in creating tools to facilitate the access to administrative data. Tools like these- the oldest EDI and the newest XBRL - represent another way of collecting data from public institutions as well as from enterprises, since they are based on the exchange of information among the data provider and the NSI on the base of a common and agreed structured data model.

Data collection process is not only a matter of interviewing techniques, but also of contact strategies as well as of monitoring activities: the first set of activities is necessary to get in touch with respondents and may vary according to the type of respondent unit (large or small enterprise, new enterprise, etc.). The second set of activities is important to keep under control the data collection while it is in progress and to take proper actions to improve or modify any factors that may badly interfere with data quality.

At the end of the data collection phase, information is ready to enter the next phase of the survey process, represented by the "*Phase 5.Process*" of the GSBPM, when data records are cleaned and prepared for the analysis. The way the following steps are faced and performed depends on how data collection is finalised since this depends on the mode(s) used to collect information.

2. General description

The data collection phase described in this module covers different sub-phases of the GSBPM that go from the design through the finalisation of data collection[1]. In more detail, readers are guided through the following steps:

- design phase
- contact and reminders strategies
- preparation activities
- collection phase
- monitoring phase
- finalisation phase

These steps are deeply described in the theme modules that are linked to the present one. Specifically, they are:

1) "Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method";
2) "Data Collection – Design of Data Collection Part 2: Contact Strategies";
3) "Data Collection – Mixed Mode Data Collection";
4) "Data Collection – Techniques and Tools";
5) "Data Collection – CATI Allocation".

The first step of the data collection phase is the design of data collection methodology.

In this step researchers determine which are the most appropriate data collection method(s) and instrument(s)[2] as well as which is the most efficient contact strategy, where efficiency is in terms of many factors such as response rate, response burden, budget constraints, etc.

How to design the data collection methodology can be found in the theme modules mentioned above: the first one ("Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method") describes which factors have to be considered when choosing data collection methods, advantages and disadvantages of each single possible mode to collect information and how these can be combined with budget and organisational constrains. The design of data collection methodology is in close connection with the questionnaire design phase, because the choice of the data collection technique is influenced by the questionnaire design and vice-versa: the various techniques allows for different interview lengths, different question formats, different question contents, different sets of checking rules, etc. This two-way influence is greater in mixed mode surveys where the questionnaire design has to take into account the presence of more techniques especially in the case when they are used concurrently.

---

[1] In the GSBPM these phases are labelled as: *2.3 Design data collection methodology*, *3.1 Build data collection instrument*, *4.2 Set up collection*, *4.3 Run collection* and *4.4 Finalize collection.*

[2] Descriptions of the various steps are derived from GSBPM and are adapted to the aims of this topic.

The design of a mixed mode strategy is treated in the theme module "Data Collection – Mixed Mode Data Collection", where both parallel and sequential mixed modes are described together with the steps to follow during the survey process (from designing to conducting a business survey) to prevent data from mode effect. Besides examples of recent mixed mode designs in business surveys from NSIs are described. They provide evidence on how to get a high response rate by using specific mixed mode strategies with unaffected overall response rates and data quality.

After the appropriate data collection mode(s) has been chosen, researchers have to design and set up the appropriate contact strategy, that is when and how respondents are contacted and what material (questionnaire, cover letter, instructions etc.) is used in each contact. In the theme module "Data Collection – Design of Data Collection Part 2: Contact Strategies" readers can find recommendations and suggestions on how to design this delicate phase of the survey process. In particular it describes which factors have to be considered, how contact strategy varies according to the type of businesses, how reminder strategy can be tuned according to the chosen contact strategy[3].

Besides, a hint is given to responsive design approach to be used for both the design and contact strategy phases and how they can be modified during the data collection process to improve response rate.

After the design phase, data collection instruments have to be built following the specifications generated during the previous design phase (Sub-phase *"3.1 Build data collection instrument"* of GSBPM). This means that, depending on the type of mode(s) used, one or more data collection instruments have to be built (paper or electronic questionnaires, SDMX hubs, systems to extract and receive data from administrative archives) and their contents and functioning have to be tested. During the building phase it is also extremely important to establish a connection between the collection instruments and the metadata system, in order to facilitate data comparability inside the entire collection system and to reduce the work in subsequent phases. Preparing also for collecting paradata will be of great help in improving the collection step (Kreuter, 2013).

After the building phase, the collection of data can start (Phase *"4.Collect"* of GSBPM). How collection is performed depends on the chosen technique. Anyway, a common set of steps to be followed in order to gather data and to get them ready to enter the subsequent phase (Phase *"5.Process"* of GSBPM) of the survey process, can be described for any data collection mode. These steps are:

- preparation activities
- collection phase
- monitoring phase
- finalisation phase

Preparation activities are those activities to be carried out in order to be ready to collect data (sub-process "*4.2 Set-up collection*" of GSBPM). They include:

- training collection staff;

---

[3] According to the GSBPM, "Contact strategy" and "Reminder strategy" are steps of the survey process that are carried out during the set up and running of the data collection phase (respectively sub-processes 4.2 and 4.3).

- ensuring collection resources are available, e.g., laptops;
- configuring collection systems to request and receive the data;
- ensuring the security of data to be collected;
- preparing collection instruments (e.g., printing questionnaires, pre-filling them with existing data, loading questionnaires and data onto interviewers' computers etc.).

The set of preparation activities can vary according to the chosen techniques: training of collection staff, for example, plays a fundamental role for interviewer-administered modes since it has to make interviewers able to collect data in the most objective way in order to reduce as much as possible the interviewer effect, that represents the effects on respondents' answers deriving from the different ways that interviewers administer the same survey (SDMX, 2009). On the other side, activities like ensuring data transmission security or availability of resources, like laptops, are peculiar of computer assisted data collection techniques.

The collection of data is run with the different collection instruments used to collect the data. It includes the initial contact with respondents and any subsequent follow-up or reminder actions. It records when and how respondents are contacted and whether they have responded (sub-phase "*4.3 Run Collection*" of GSBPM). For CATI surveys, the management of contacts with respondents is described in the theme module "Data Collection – CATI Allocation" that focuses on this peculiar feature of CATI, represented by the scheduling of telephone calls among the interviewers.

The monitoring phase is run while data collection is in progress in order to allow researchers to keep it under a constant control. Monitoring is based on a set of indicators about different aspects of the data collection like interviewers' productivity, response rate, non–response rate, refusal rate, interview length etc. In general, a unified system of codes for each indicator, to be used for any business surveys run inside an NSI, would be of a great help in computing comparable indicators (Györki, 2012). Besides, it would be advisable to use these codes to build quality indicators (see also the module "Quality Aspects – Quality of Statistics") that will help monitoring the different problems that might arise during data collection that can cause non response errors, coverage errors and measurement errors (Eurostat, 2009). Example of these problems are:

- frame problems (status of the statistical unit: dead, under liquidation, etc.), classification problems, accessibility problems;
- problems with the activity of the statistical unit (no business activity: now, never, temporarily)
- problems referred directly to respondent (refuse to provide cooperation, no successful contact with him, etc.).

Finalisation of data collection starts when the collection of data is over. This step includes loading the collected data into a suitable electronic environment for further processing (4.4. Finalize collection – GSBPM). How finalisation of data collection is performed strictly depends of the technique used, being the computer assisted ones able to facilitate and speed it up. In fact, for these techniques, data are already stored in an electronic format and (partially) checked during the data collection itself. The consistency of final data can be further improved if the survey questionnaire has been designed following a metadata-driven approach or any techniques for relational database design.

How the above steps, from the preparation activities to the finalisation of data collection, have to be managed for the various data collection techniques is described in the theme module “Data Collection – Techniques and Tools”. This module considers only the main collection modes and divides them in two groups: interviewer-administered and self-administered. The first includes CATI, CAPI and Direct Observation while the second contains Mail and Web surveys. Besides, administrative data as well as data transfer through EDI and XBRL are also described.

How to collect statistical information from other data sources different from surveys is described in the module “Data Collection – Collection and Use of Secondary Data”. The entire process of collecting already existing data is generally referred to as the collection of secondary data. The topic discusses the advantages and disadvantages of this approach from an official statistics point of view together with research strategies with secondary data. A classification of secondary data types and an overview on the different types of use of secondary data by NSIs can also be found in this topic.

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

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

GSBPM (2009), Generic Statistical Business Process Model. Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS) Version 4.0 – April 2009.

Györki, I. (2012), GÉSA: The Tool for Survey Control, Quality Assessment and Data Integration. *Hungarian Statistical Review, Special number* **15**, 48–78. http://www.ksh.hu/statszemle_archive/2012/2012_K15/2012_K15_048.pdf

Kreuter, F. (2013), *Improving Surveys with Paradata: Analytic Uses of Process Information*. Wiley.

SDMX (2009), Content-Oriented Guidelines Annex 4: Metadata Common Vocabulary 2009.

Interconnections with other modules

8. Related themes described in other modules

1. Data Collection – Design of Data Collection Part 1: Choosing the Appropriate Data Collection Method
2. Data Collection – Design of Data Collection Part 2: Contact Strategies
3. Data Collection – Mixed Mode Data Collection
4. Data Collection – Techniques and Tools
5. Data Collection – CATI Allocation
6. Data Collection – Collection and Use of Secondary Data
7. 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.

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

Data Collection-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-02-2012	first draft	M. Murgia	ISTAT (Italy)
0.2	03-08-2012	second draft	M. Murgia	ISTAT (Italy)
0.3	05-09-2012	third draft	M. Murgia	ISTAT (Italy)
0.4	19-11-2013	fourth version after EB revision	M. Murgia	ISTAT (Italy)
0.4.1	21-11-2013	preliminary release		
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 17:48