Short paper

Using survey data collection as a tool for improving the survey process

Silvia Biffignandi (University of Bergamo), Antonio Laureti (Istat), Giulio Perani (Istat)

Abstract.

The paper focuses on how paradata information can improve survey methodology and quality.

Collecting data via Web allows for server-side paradata (i.e. log files describing access time, number of accesses and so on) and client-side paradata (namely the answering process within the questionnaire, i.e., insight into the sequencing and completeness of responses. Respondent behaviour is traced on each Web page as they answer the survey). If we consider the business surveys, business register data are available, too. This data may be linked to survey data. Therefore, an integrated set of data becomes available and may be used not only for descriptive purposes of substantive information, but for improving the data collection process at different steps of the survey.

This paper is analyzing data collected using a web questionnaire in Italy (Research and Development, R&D, Business Survey, carried out from Istat) and discusses how survey response, client-side paradata, survey-side paradata and auxiliary variables from the business register can be allocated in the framework of the survey process.

1. **The problem**

Web survey mode allows for the collection of paradata during web questionnaire completion; these data that are generated during the fieldwork of the survey (Biffignandi and Bethlehem, 2011). We can distinguish between server-side paradata and client-side paradata (Biffignandi S. 2010). Server-side paradata are collected by software tools running at the server where the questionnaire is located. They relate mainly to the questionnaire compilation process, like the number of times the questionnaire is accessed, the time spent at each access, the type of browser used and so on. This data are contained in the so called logfiles. Client-side paradata describe how the respondents are answering the questions (order, questions skipped, keys that have been pressed and so on).

We analyse paradata; these data could be used for several purposes: from the identification of the most difficult questions in the survey form, to the identification of some missing (or too restrictive) checks, from the degree of ready availability of R&D data in enterprises, to the time needed to fill in the questionnaire (and, in turn, to quantify the burden on respondents). Our analyses may provide insight in how easy or difficult it is for the respondents to complete the questionnaire and how to improve it, if there are some sectoral activities for which the survey is particularly difficult or burdensome.

2. **The survey and the data collection**

The Italian Statistical Office (Istat) is collecting business R&D data and producing official statistics on the Italian business R&D activities since 1963. The Istat business R&D survey, carried out on an annual basis, follows the methodological recommendations provided by the "Frascati Manual" (OECD, 2002), the main source of theoretical and practical guidelines to undertake statistical R&D surveying at international level[1].

On the basis of information collected by administrative sources (such as official statistical business register, fiscal data from the Italian Tax Authorities, the Italian Register of R&D performing institutions, managed by the Ministry of University and Research (Miur), data on national and EU funding to research projects, patent databases, private business reports), all enterprises known, or assumed, to be R&D performers are included in the R&D survey. There isn't a cut-off point for the enterprise size. No special treatment is implemented in data collected on micro-enterprises (0 to 9 employees) providing that they are employing at least one researcher[2]. However, for selected industries and technological areas, also micro-enterprises with less than one researcher are included in the realised sample.

Around 20,000 enterprises are currently under monitoring as "potential R&D performers" in Italy. Most of these enterprises are regularly surveyed to identify the "actual R&D performers". Overall response rate is around 55 per cent with reference to the year 2008, including both R&D performers and non-performers. The evolution of the Italian R&D survey population size and related rates of response is shown in table 1 below.

Table 1. Italian R&D survey: target population and response rate. Years 2000-2008

Reference year of the survey	Target population (number of enterprises)	Number of enterprises reporting actual or planned R&D activities	Overall response rate (percentage of the target population)
2000	16,294	2,367	49.0
2001	15,377	2,684	61.4
2002	26,149	3,222	55.6
2003	24,708	3,345	49.7
2004	19,962	3,457	42.7
2005	24,914	3,384	36.3
2006	26,237	4,419	43.5
2007	16,730	4,850	51.4
2008	17,631	6,088 (including 5,467 actual R&D performers)	54.7

As to the data collection method implemented by Istat, it can be stressed that the R&D survey has probably been the last Istat business survey to adopt an electronic questionnaire. The main reason for relying on the traditional paper forms has been that the questionnaire is very complex. There is a close interdependency among a number of questions included in the questionnaire and there is need for allowing the respondents to easily come back on specific questions to make them consistent with data provided in other sections of the same questionnaire. Implementation test (undertaken in the 2006 survey) of an "off-line" electronic questionnaire (MS Excel datasheet) was unsatisfactory. In order to overcome problems, an advanced

[1] The "Frascati Manual. Proposed Standard Practice for Surveys on Research and Experimental Development" is an OECD technical manual, firstly published in 1963, aimed at providing the OECD countries with common methodologies to estimate their own investments on R&D. The Manual, now available in its sixth edition, is currently used all over the world and is considered – according to the EU legislation – as the main methodological source for the production of official R&D statistics in the European Union as well.

[2] Usually expressed in "full time equivalent" (FTE), as recommended by the Frascati Manual.

design for a Web-based electronic R&D questionnaire was developed in 2008 in order to be implemented already in the data collection round with reference to the year 2007.

The structure itself of the questionnaire of the Istat Business R&D survey is firstly aimed at isolating R&D performers from non-R&D performers. At the beginning of the questionnaire, a filtering question leads the respondents to one of the routes allowed to respondents who could be broken down into three main groups: actual R&D performers (those who have to provide extensive information on the R&D activities undertaken in the reference year); future R&D performers (who are just asked to report about their future investment plans) and non-R&D performers (no data requested). Of course, any reference to the complexity of the questionnaire and the associated burden for respondents is relevant only to actual R&D performers. Beyond the filtering module, the questionnaire contains 24 questions which belong to three main areas: questions on R&D expenditure (quantitative), questions on R&D personnel (quantitative) and qualitative questions on the R&D projects undertaken by the enterprise. Most of the consistency checks performed by the data collection tool refer to the internal consistency (at least in terms of totals) among the questions on R&D expenditure and cross-checking between R&D expenditure and R&D personnel data to assure a logical consistency between them.

The structure of the Web questionnaire was largely based on the previously used paper forms but the overall architecture was very innovative by unbundling, on the one hand, the “electronic form” – just a basic frame to be easily administered in a Web environment – and, on the other hand, its “smart” component containing a tool for the identification of the respondent and the correct sequencing of the questionnaire’s provision, as well as including more than 400 checking rules. These two components – electronic form and checking tool - are physically distinct, being, the first, delivered via Web on the remote PC of the respondent and the, second, resident in a Web-server in the Istat’s premises.

Besides the regular assessment on the effects of the implementation of a new data collection tool on the data production processes at Istat (by considering indicators such as the overall length of the data collection process, the time needed to check the collected questionnaires for errors and inconsistencies, the feedback by respondents, etc.), the software implemented at Istat allows for producing a set of information on the behaviour of the survey’s respondents .

3. Paradata

A few key concepts have to be clarified before discussing before presenting the available paradata. Those of “event” (with the associated timing), “access” and “error”.

All the available paradata are based on a series of “events”. We can identify (and record) an “event” each time a respondents is interacting with the electronic form (basically, by typing a figure or a word, or even, by scrolling down the form itself). We can identify each single “event” by its nature and by the time when it took place, as well as by its duration in time (usually, from a fraction to a few seconds). Each event – or a group of events – is associated to an “access”.

The “access” itself to the questionnaire cannot be identified in a straightforward way as “log-ins” and “log-outs” are not recorded in the system. We know that a respondents is (was) connected to the server because of the activity carried out on the questionnaire. As a consequence, an access without any activity on the questionnaire will not be taken into consideration in this analysis. Moreover, an access has to be qualified in terms of time. Conventionally, all events taking place within one hour (or three, or six hours) could form an “access”. For the purposes of this study, only “daily accesses” have been taken into consideration, i.e. an “access” will be equivalent to the set of all events having taken place in a day.

Finally, it should be pointed out that events could be either “correct” (according to the rationale behind the structure of the questionnaire) or leading to the generation of “errors”, mainly inconsistencies in relation with other pieces of information previously provided through the questionnaire. It is obvious that, as the questionnaire is “completed” only when all “errors” are properly fixed, the generation of “errors” will have as a result the need for new events aimed at correcting them. In the process of numbering the errors produced by each respondent, only the first appearance of an error type was taken into consideration, even in the event that the same error type would have been repeated in more than one session.

Using “event” (with the associated timing), “access” and “error” concepts is possible to compute some basic indicators (see table 2). Computing the indicators on the 2008 Istat Business R&D survey, the “actual R&D performers” appear to be by far the most interesting group, having intensively used the data collection tool (and experienced most of the advantages/problems to use it, at least in terms of dealing with the consistency checking structure).

To complete a questionnaire, an enterprise – on average – needed to type 236 digits by accessing it during 4.5 daily sessions and producing 19 errors.

In addition to this set of basic indicators, several information on the use of the data collection tool is actually available, as to the timing of use, intensity of use, compilation routes and generation of errors:

- Timing of use:
    - Entry and exit date.
    - Days in which the tool has been accessed.
    - Hours of use during the day.
- Intensity of use:
    - Number and typology of actions (events).
    - Processing time associated to each single action.
- Compilation routes:
    - Outcome of the compilation process.
    - Sequencing in accessing the questions.
    - Changes to previously compiled questions.
- Generation of errors:
    - Number and typology of errors.
    - Questions (or groups of questions) mostly affected by errors.

All these information, being referred to single respondents, can be also analysed by considering some key features of the concerned enterprises, such as number of employees, economic activity (in terms of NACE classification), geographical localisation (for instance, in terms of NUTS2 regional classification) and possible position within an enterprise group.

Most interesting paradata are breaken-down by stratification variables, such as size, sectoral activity, enterprise group belonging.

4283 out of 6088 respondents were participating to the survey via web. 4136 were participating without errors. The results are discussed taking into account the complex structure of the questionnaire

Table 2. Business R&D survey 2008 data collection. Paradata indicators: Number of respondents*, daily accesses and errors.					
Paradata indicators	Actual R&D performers	Future R&D performers	Non R&D performers	Respondents who have accessed the questionnaire without completing it	Total
Number of respondents*	2.601	268	1.086	324	4.279
Number of events	614.273	10.942	30.577	5.274	661.066
Average number of events per respondent	236,17	40,83	28,16	16,28	154,49
Number of daily accesses	4.471	371	1.359	391	6.592
Average number of accesses per respondent	1,72	1,38	1,25	1,21	1,54
Average number of events per access	137,39	29,49	22,50	13,49	100,28
Number of errors**	48.291	860	1.852	468	51.471
Average number of errors per respondent	18,57	3,21	1,71	1,44	12,03

* Data available only for the respondents who filled in the electronic questionnaire.

** The errors refer to a total of 4,136 respondents, as 147 of them apparently made no errors in filling in the questionnaire.

5. References

Biffignandi S. & Bethlehem J., Handbook on Web surveys, Wiley, N.Y. 2011.

Biffignandi S. (2010), Modeling non-sampling errors and participation in Web surveys, (invited speech), *Proceedings of the 45th SIS Scientific Meeting*, Padova, June 2010, ISDN 978-88-6129-566-7.