Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

DATA COLLECTION CHALLENGES AT STATISTICS NETHERLANDS

Jelke Bethlehem[1]

ABSTRACT

Statistics Netherlands is confronted with several developments in society that have a substantial impact on its task of collecting, processing and publishing statistics. Most importantly, Statistics Netherlands has to reduce the administrative burden put upon companies and households. If relevant data are available elsewhere, they should not be collected once again in a survey. This means a shift in focus. Wherever possible, data available in existing registers should be used. Only if such data are not available, surveys can be conducted to collect them. And preferably this is electronic data collection. This paper is about some of the challenges making this shift.

KEY WORDS: Registers, Web –surveys, Primary data collection, Secondary data collection

1. INTRODUCTION

1.1 The problem

Statistics Netherlands is confronted with several developments in society that have a substantial impact on its task of collecting, processing and publishing statistics. Most importantly, Statistics Netherlands has to reduce the administrative burden put upon companies and households. If relevant data are available elsewhere, they should not be collected once again in a survey.

With respect to published statistics, Statistics Netherlands is confronted with some dissatisfaction with respect to their usefulness. Customer demands are changing, and also new customers appear on the market. There seems to be a growing demand for more thematic publications (in which data from various sources are combined to provide an integrated view), and for statistics on detailed regional levels. Also, customers demand more timely data.

Statistics Netherlands faces the problem of taking up these challenges while its available budget is constantly being reduced.

1.2 A shift in focus

With regard to data collection, these developments mean a shift in focus in two ways:

- A shift from primary data collection to secondary data collection. Wherever possible, data available in existing registers should be used. Only if such data are not available, or the available data is of insufficient quality, timeliness, or coverage, surveys can be conducted to collect them.

- A shift from more traditional ways of primary data collection to new cheaper and faster ones. Such surveys should preferably be carried out in electronic ways, i.e. using some mode of computer-assisted data collection including Internet surveys.

To meet these challenges satisfactorily, a substantial transformation appears to be necessary. Changing from survey-based statistics to register-based statistics has a substantial impact on the organisation. The combination of various registers in an integrated system requires a new data infra-structure, new metadata systems, new quality control

systems, and new register methodology to ensure sufficient data quality. Moreover, it needs a new way of thinking. Also, it should be explored whether Internet surveys can fulfil the need for additional data. This paper describes some of the challenges of this transformation process.

2. SECONDARY DATA COLLECTION

2.1 Changing to register-based statistics

Although Statistics Netherlands only generates a small part of the total response burden caused by the Dutch government (0.03%), it experiences an ever increasing demand for a substantial reduction of the administrative burden on businesses and households. Statistics Netherlands has already been very successful in reducing this response burden, but still a lot of work can be done. One way to realise this is a change from survey-based statistics to multi-source register-based statistics. This is made possible by a new statistics act, which gives Statistics Netgherlands access to official registers. This change means

- Focus on secondary data collection, i.e. using data available in existing registers where possible;
- Primary data collection, i.e. conducting surveys only when required data are not available in existing registers;
- Conducting remaining surveys using computer-assisted data collection instruments (including Internet surveys);
- Creating an integrated data-infrastructure, within which all available data and metadata are stored in such a way that the information in the statistical databases is interlinkable;
- Efficient production methods by standardising concepts, methods and tools.

Most registers are *administrative registers*. Registers can also be used for statistical purposes. This can take several forms. The first form is to use a register as a primary source of data for statistical analysis. If the register contains the right variables, their values are available for every element in the population (or sub-population) covered by the register. Population quantities can be computed directly, and without uncertainty due to sampling variance. An advantage of direct use of register data is that they can be used for computing statistics on small areas.

A typical example of such use of a register in The Netherlands is the GBA (Gemeentelijke Basis Administratie voor persoonsgegevens). It was introduced in 1994. It is a comprehensive and cohesive registration system for population data. It is fully decentralised. Every municipality has its own population register containing basic data on all its inhabitants. Statistics Netherlands uses this register for compiling demographic statistics. These data are also used to construct models for population forecasts.

A second form of use of a register is as a *sampling frame*. It is a list that unambiguously identifies every object in the population to be investigated. It is used as the basis for selecting a sample from the population. The information in the register should be such that every selected object can be located and contacted. This means address, telephone number, e-mail address, or other contact information should be available.

Statistics Netherlands uses the GBA as a sampling frame for its social surveys. An example is the Integrated Survey on Living Conditions. It is a large continuous survey. Every month a sample is selected. Persons are selected by means of a stratified two-stage sample from the GBA. In the first stage, municipalities are selected within regional strata with probabilities proportional to the number of inhabitants. In the second stage, an equal probability sample is drawn in each selected municipality.

A third form of use of registers is for *weighting adjustment*. Many surveys are affected by nonresponse. If nonresponse leads to biased estimates, wrong conclusions are drawn from the survey results. To avoid this, some kind of correction procedure must be carried. One of the most important correction techniques for non-response is *adjustment weighting*. It means that every observed object in the survey is assigned a weight, and estimates of population characteristics are obtained by processing weighted observations instead of the observations themselves.

As an example, a number of variables that are available in the GBA (sex, age, marital status and region) are used for weighting adjustment in the Integrated Survey on Living Conditions

2.2. Quality of register data

Statistical use of register data is a means of secondary data collection. Data are already available in electronic form. Therefore there is no response burden, and data collection costs are low compared to surveys. However, it should be taken into account that data have been collected by different agencies for different purposes at different times. This gives rise to the question whether register data are of the same quality as survey data.

In survey-based statistics, estimates are computed based on a sample from the target population (the population to be investigated). Such estimates will never be exactly equal to the population characteristics to be estimated. There will always be some error. This error can have many causes. Two broad categories can be distinguished: sampling errors and non-sampling errors.

Sampling errors are substantially reduced by use of registers. In most cases the sample size is very large, if not equal to the size of the complete population. In the latter case, sampling errors vanish completely. Non-sampling errors are not automatically reduced if register data are used. Substantial errors may still be caused by phenomena like item non-response, measurement errors, and lack of population coverage. Some basic registers are of high quality data, but in other registers data quality may be affected by non-sampling errors. For example, the population register (maintained by municipalities) and business registers (originating from chambers of commerce and tax data) suffer from both over- and under-coverage due to the dynamics of these populations.

Probably one of the most important differences between registers and surveys is that generally the amount of unit non-response is small. Unit non-response plays an important role in surveys, especially when it is caused by the topic of the survey. This seems to be far less the case for registers, because participation in the data collection is often obligatory or beneficiary to the data providers. However, some nonresponse problems will always be present.

Many registers that are currently used by Statistics Netherlands need to be edited and imputed before they can be used for statistics. The degree to which these registers contain missing data and measurement errors differs for each register and is related to the objectives of the register holders and the extent to which register holders monitor the quality of their data. In this respect there is no difference in quality compared to surveys.

Given the fact the register data are not of perfect quality, the next question is to what extent Statistics Netherlands should monitor and improve the quality of register data. And if some quality control system is needed, then the question is whether to use the quality control system of the register owner (if it exist at all), or to develop a new system that satisfies the requirements of Statistics Netherlands? It is current practice at Statistics Netherlands to monitor the quality of the register data it receives. The administrative register data are processed and transformed, and so a statistical register is formed. However, due to the diverse nature of the registers, currently there are no general methodological framework and tools for quality control systems. And it is not an easy task to develop such a system either, since the contents, formats and possible sources of error are very diverse. As a consequence the quality level of registers can be unclear.

Development of a quality control system cannot be realised without register holders. They are responsible for the data collection and processing. They will perform their own quality checks based on possible sources of errors that they have identified. For instance, the digital tax forms issued by the tax board contain many edit rules similar to rules implemented in Blaise CAPI and CATI questionnaires developed by Statistics Netherlands. In practice Statistics Netherlands is often a small player with relatively little influence on the process of data collection and processing performed by the register holder. The reason is that most registers are collected for different purposes than making statistics.

A good example is the new data collection process on jobs and allowances data that will start in 2006 and that will be fully under the responsibility of the tax board and the social security board. At the same time, Statistics Netherlands will discontinue the additional survey on employment and wages (Enquête Werkgelegenheid en Lonen). These changes have considerable impact on the statistical process of Statistics Netherlands. Unfortunately, it took Statistics Netherlands substantial efforts to be recognised as a party in the transformation process.

In conclusion, it can be said that quality control exists at least to some extent, but register holders and Statistics Netherlands have different demands and systems. For efficient future data quality management it is important to develop a general quality framework for register data, based on a classification of sources of errors and possible treatments. Ideally, this should be done in close co-operation between Statistics Netherlands and register holders. In order to be able to do this the position and role of Statistics Netherlands should be recognized and formalised.

2.4. A register model

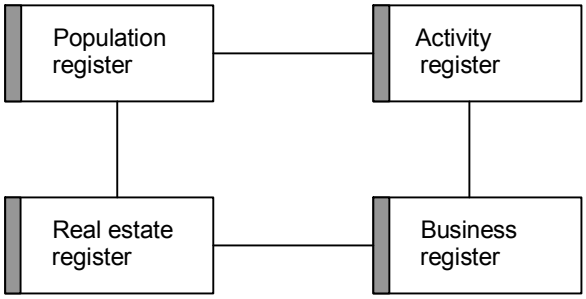
In principle, registers are nothing more than big survey data files. Therefore, register data can be processed in much the same way as survey data, using familiar editing and imputation techniques. What is new, however, is the integration of registers with other registers and with surveys. There are a number of challenges involved in this new approach.

First, a thorough inventory has to be made of registers that exist and can be useful for register-based statistics. Currently, it is not clear whether Statistics Netherlands has a complete overview of all relevant existing registers.

Second, there are conceptual challenges. Objects and variables measured in registers do not always correspond to statistical objects and variables. And if they do not correspond to each other, is it possible to transform one into the other? And if not, should Statistics Netherlands change it definitions, thus causing breaks in time series?

Third, new estimation theory should be developed for statistics based on combination of data from several sources (both registers and surveys). The new theory must be implemented in a workable environment. This may also mean a shift in thinking. On the one hand, maintaining administrative registers is a form of book-keeping. It means that individual records have to be correct. On the other hand, statistical use of registers means that accurate estimates must be computed, and this not necessarily means that each individual record is without error.

Figure 2.4.1. A model for register statistics



Use of an integrated system of registers and surveys offers great opportunities to produce integrated statistics and makes it possible to answer new research and policy questions. Statistics Sweden had made a model for an integrated system of registers, see Statistics Sweden (2001). In a simplified form, this model is displayed in figure 1. A similar model may be developed in Statistics Netherlands.

Crucial for a system of interlinked topical registers is that all elements in all registers can be uniquely identified. This requires a system of identification numbers. Partly such a system already exists, or will be developed in the near future; partly the situation still is unclear, depending on a government-wide initiative for “streamlined base registrations (SBG)”. Only when all information from various registers can be properly linked, will Statistics Netherlands be able to fully take advantage of the available data, and to produce relevant, comprehensive thematic statistical publications.

Another important aspect of such a system of registers is that all data are properly documented. This requires comprehensive, standardised system of metadata. Systems should be able to automatically generate documentation from the metadata. Also, process metadata should be available in electronic way in such a form that it can be used as input for data processing systems.

Registers are usually updated on a regular basis. This means that new versions of data sets will become available frequently. This requires an adequate and efficient version control system.

3. PRIMARY DATA COLLECTION

3.1 Internet surveys

The existing registers in The Netherlands are not sufficient to fulfil the need for statistical information about Dutch society. Therefore, additional surveys are still needed. To reduce the costs of primary data collection, Statistics Netherlands investigates the possibility of replacing its traditional paper questionnaires (PAPI), CATI and CAPI surveys by Internet surveys.

At first sight, Internet surveys seem to have some attractive advantages. First, it is a simple means to get access to a large group of potential respondents. Second, no interviewers are needed, and there are no mailing costs. Third, surveys can be launched very quickly. No time is lost between the moment the questionnaire is ready and the start of the fieldwork. And fourth, particularly web surveys offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation, and movies). For these reasons, Internet surveys have grown in popularity. Particularly, commercial organisations like Harris Online, Bloomerce and McKinsey have implemented large web panels at an international scale. However, use of web surveys is not without methodological problems. These are partly caused by using the Internet for selecting respondents, and partly by using the web as a measuring instrument. If these problems are not seriously addressed, Internet surveys may result in low quality data for which no proper inference can be made with respect to the target population of the survey.

3.2. Using the Internet for economic surveys

Most experiments of Statistics Netherlands with web surveys have been carried out in the area of business surveys. This is because penetration of computers and Internet facilities is much bigger in companies than in households. The first Internet survey test of Statistics Netherlands was carried out in 1998. It was investigated to what extent e-mail could be used to collect data for the Business Tendency Survey. The test was a success. The response rate among the participating companies was almost 90%. Overall, respondents were positive. However, they considered the simple ASCII-based questionnaire text old-fashioned, and not very user-friendly.

Following the success of this test, it was decided in 1999 to carry out a test with Blaise IS, the first version of the Blaise System capable of generating web questionnaire forms. The target population was a group of approximately 1500 construction companies. The survey form for the monthly statistics on construction was used. It contains only three questions related to turnover and number of employees. Companies could choose between three modes of data collection: an e-mail-survey, an on-line web survey and an off-line web survey. The overall conclusion of this and the previous test was that this way of data collection is certainly feasible, and that in course of time it could make data collection more efficient and cost-effective.

The success of these two tests leads to the *E-Quest Project*. Objective of this project was to implement the use of Internet survey forms in the production environment for some 20 short-term panel business surveys. Another objective of the project was to convince three out of four of the 27,000 companies in the panel to participate in this form of electronic data transfer. After intensive efforts during the past years, now about 18,000 companies receive monthly or quarterly an electronic questionnaire. About 75% of the large companies respond electronically. Some companies cannot surf on the web. These e-mail-only companies are sent (by e-mail) a monthly or quarterly questionnaire that can be opened and handled with the special EDR-software (see below). The web-surfers are sent an e-mail with an HTML-form as attachment. Data encryption is used to keep the data confidential.
In October 2004, a pilot has started to find out whether electronic questionnaires can be used in data collection for the Annual Business Inquiry. Two approaches will be tested. The first approach is an off-line method, denoted by Electronic Data Reporting (EDR). This system helps respondents to manage Blaise interviewing programs on their own computers. After the EDR-software has been installed on the computer of a respondent, new survey interviews

can be sent by e-mail. They will be automatically imported in the EDR environment. A simple click will start the interview. After off-line completion of the interview, the entered data are automatically encrypted and sent to Statistics Netherlands. The second approach is an on-line web survey. The survey questionnaire is generated from the same Blaise data model. Respondents are allowed to interrupt the interview, save the entered data, and continue at a later point in time. The pilot will be carried out in a sample of 100 companies in four different branches: transport of goods by road, ICT-companies, accountancy firms, and professional cleaners.

The EDR approach is presently also in production with the new survey on international trade in services. A small but increasing percentage (below 10%) has difficulties caused by firewalls, no-foreign-software policies and dependence on external ICT support. EDR is also used for road transport statistics. Finally, an advanced off-line Blaise questionnaire for the Survey on producer price indices has been developed and tested. It is scheduled to be in production at the end of 2005.

These experiments have made clear that there is not one optimal mode of data collection that can be used for all participants in business surveys. Therefore, there is a need for mixed-mode data collection. This calls for a proper environment capable of generating questionnaires for different modes from a single questionnaire specification. Questionnaires should be optimally adapted for the mode in which they are used. But also mode effects should be avoided as much as possible. Vital for a mixed-mode approach is the implementation of a good case management system allowing, for example, cases to be moved from one mode to another in an efficient way without losing control.

To co-ordinate, manage and control the flow of electronic questionnaires from businesses to Statistics Netherlands, a new organisational unit will be formed: the Data Contact Centre (DCC). This unit will be responsible for all technical aspects of electronic data collection, like sending questionnaires, receiving data, sending reminders, case management, security, data storage, and supplying subject-matter departments of Statistics Netherlands with the proper data.

3.3 Using the Internet for social surveys

In their fundament paper, Horvitz and Thompson (1952) show that only unbiased estimates of population characteristics can be computed if every element in the population has a non-zero probability of selection, and these probabilities are known to the researcher. Furthermore, only under these conditions, the accuracy of estimates can be computed. Many Internet surveys are self-selected surveys. Open invitations on portals, frequently visited websites, or dedicated survey sites are used to select respondents. The probability of being confronted with such an invitation is unknown as well as the probability of accepting it. Therefore, it is impossible to compute unbiased estimates of population characteristics.

A probability sample is nearly always selected using a sampling frame. Problems can arise when the population represented by the frame differs from the target population of the survey. *Under-coverage* occurs when elements of the target population do not appear in the sampling frame. This can be a serious problem for web surveys. Usually, the target population of NSI surveys is much wider than just those people with an Internet connection. Therefore, a substantial part of the population will never be selected if the Internet is used as a sampling frame. Indeed, many people do not have access to the Internet (yet). As can be seen in figures 2 and 3, Internet use in the Netherlands decreases with age, and increases with the level education (source: Statistics Netherlands, 2003).

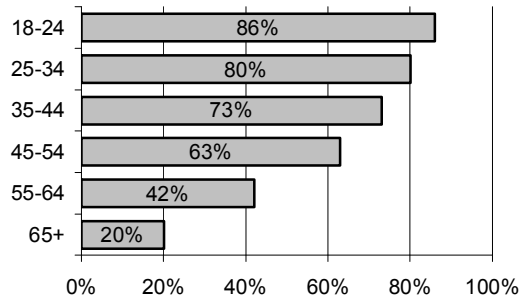


Figure 2. Internet use by age. Source: CBS

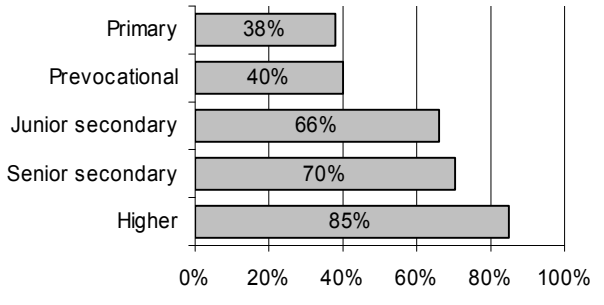


Figure 3. Internet use by education. Source: CBS

Nonresponse can also be a source of serious problems. An Internet survey questionnaire is a self-administered questionnaire. Therefore, nonresponse rates may be high for this type of survey. Technical problems may be an additional source of nonresponse problems. Slow modem speeds, unreliable connections, high connection costs, low-end browsers, and unclear navigation instructions may frustrate respondents. In order to keep the survey response up to an acceptable level, every measure must be taken to avoid these problems. This requires a careful design of web survey questionnaire instruments.

One way of using web surveys for data collection is to let it be one of the modes in a mixed-mode survey. By combining various modes (e.g. web, telephone and face-to-face), it may be possible to achieve high response rates and high data quality at low costs. In designing and implementing a mixed-mode survey, a survey organisation is confronted with a number of challenges. First, to avoid mode effects as much as possible, the questionnaire forms used in different modes should resemble each other as much as possible. Second, often different systems are used for different modes. Then it will not be easy to design, implement and maintain data collection instruments that work exactly the same on different systems. Mode effects can be introduced very easily, especially for less concrete and more subjective issues. Third, if different modes are carried out with different systems, case management may prove to be complex. And fourth, if different systems are used, data may be stored in different files formats at different locations. Merging of various data files into one analysis data file may turn out to be cumbersome.

In order to avoid the problems mentioned above, there is a need for a well-integrated data collection system. Such a system should not only support traditional modes of data collection, but also data collection over the Internet. Already some years ago, Statistics Netherlands has developed the idea of a Control Centre for Computer Assisted Survey Processing, see Bethlehem (1997). It has been implemented in a software system called *Blaise*. To carry out a survey, Blaise requires the definition of a data model. It is a description of the questions to be asked, and the rules the answer have to obey. Blaise stores this data model in machine-readable form. Blaise supports several modes of data collection: traditional data collection with PAPI, CAPI, CATI and CASI. The current version of Blaise also supports computer-assisted web interviewing (CAWI). From the same data model, a web survey questionnaire can be generated. Two approaches to web interviewing can be implemented. The *question-oriented approach* is used for long and complex questionnaires with routing and checks. The respondent remains on-line during the entire interview, and sees one question at the time. The *form-based approach* is used for short and simple surveys with straightforward data entry without question routing. The web page contains a single form with all questions. The form can be completed off-line.

Now that more and more households in The Netherlands have access to these facilities (see figure 4), it has become interesting to investigate their use for household surveys. A first step in this direction has been the *CBSquest* tool for off-line questionnaires. For this first test a question module was used from the Integrated Survey on Living Conditions (POLS). To check whether the whole procedure was technically correct, a test was carried out on a sample of 40 employees of Statistics Netherlands. This test was a success. The second step consisted of an experiment in which the non-respondents of a telephone survey on ICT use were approached again asking them to fill in an on-line web survey or a mail survey. One objective of this project was to investigate whether Internet surveys can be used for population groups that are hard to reach in telephone surveys (unlisted number, mobile phone). Another objective was to explore possible cost reductions of this means of data collection. In the same

experiment a fresh sample of 9,000 respondents was approached by advance letter asking to fill in a web survey. The response rate, after intensive follow-up, was around 30%. Analysis is still going on as to what extent the responding sample is representative. Also possible mode-effects are explored. The next step will be to develop an electronic version of the Consumer Sentiments Survey. Because of its very subjective nature this questionnaire is rather sensitive to mode-effects.

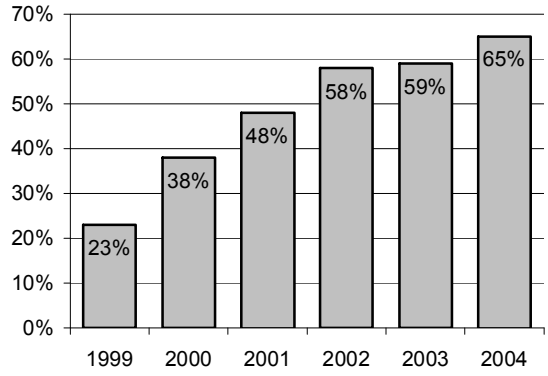


Figure 4. Households with Internet. Source: CBS

In order to cut costs of primary data collection, national statistical institutes may consider to use access panels. This is group of respondents that have been recruited to periodically fill in web questionnaire forms, for consecutive waves of the same (panel) survey, or for different surveys. Particularly, when a group of respondents has only been selected once, and can be used for several surveys, costs can be reduced substantially. For this reason, and the relative ease to get a large amount of participants, access panels have become popular among commercial market research agencies. They also claim that "on-line research is an unstoppable train".

Access panels suffer from a number of methodological problems, the most important one being lack of representativity due to under-coverage and self-selection. One way to improve representativity is to conduct a small additional control survey in the traditional way, with a real probability sample. The information collected in this way can be used in some kind of weighting adjustment procedure. Currently, it is not clear yet to what extend such control surveys are sufficient to obtain accurate, unbiased estimates. Moreover, they also increase data collection costs. There are other differences between access panels and traditional CAPI and CATI surveys. One is that there are no interviewers to assist respondents in filling in the answers to questions. Another one is that CAPI and CATI survey instruments usually have built-in checks on the consistency of answers. Confronting access panel members with error messages may easily frustrate them, causing them to stop filling in the form. As Dillman (2005) puts it: "Error messages are the death of web surveys". All this may lead to data of lesser quality. Further research should show whether web surveys and access panel may be appropriate data collection techniques for national statistical institutes.

AKNOWLEDGEMENTS

The author wishes to thank Dirkjan Beukenhorst, Johan Lammers, Frank van der Pol and Ger Snijkers for their valuable comments and contributions to this paper.

REFERENCES

Bethlehem, J.G. (1997), Integrated Control Systems for Survey Processing. In: L. Lyberg et al., *Survey Measurement and Process Control*. Wiley, New York, pp. 371-392.

Dillman, D. (2005), *How Internet Surveys are changing Data Collection Practices*. Presentation at the First EASR Conference, Barcelona.

Horvitz, D.G. and Thompson, D.J. (1952), A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, pp. 663-685.

Statistics Sweden (2001), *The Future Development of the Swedisch Register System*. R&D Report 2001:1, Statistics Sweden.