HANDBOOK
ON DESIGN AND IMPLEMENTATION
OF BUSINESS SURVEYS

Edited by

Ad Willeboordse

October 1997

PREFACE

In the last decade a strong drive took place in the European Union to come to the creation of harmonized business statistics. With the adoption in December 1996 of the Council regulation concerning structural business statistics a last building block was added to a comprehensive statistical infrastructure comprising concepts, classifications, methods and tools.

With a new legislative groundwork in the process of being implemented Eurostat felt the need to provide to European statisticians a training document on the subject. The objective of such a book would not be to provide definitions, fixed methods or rules. Rather the created "Handbook on the design and implementation of business surveys" gives in a step by step approach guidelines. A set of pragmatic ideas for making good statistics.

We think the book will support statisticians in Member States when making changes in the data collection and processing process. We hope it will assist statisticians of pre-accession countries when implementing changes in the national statistical data collection structure in order to prepare for membership of the European Union. And we are confident that it will contribute to the creation of high quality statistics for the European Union, not in the least through supporting the TES -Training for European Statisticians- program.

This manual has been created by Ad Willeboordse of the CBS - Statistics Netherlands on the request of and in close cooperation with Eurostat D2 under the responsibility of Mr. Marco Lancetti. The views in this publication do not necessarily reflect the official opinion of the European Commission.

Luxembourg, October 1997

ACKNOWLEDGMENTS

Eurostat and Mr. Willeboordse gratefully acknowledge the following co-authors for their valuable input:

- Hans Akkerboom — Statistics Netherlands
- Peter Bøegh-Nielsen — Statistics Denmark
- Cor Citteur — Statistics Netherlands
- René Huigen — Statistics Netherlands
- Karin Hilbink — Statistics Netherlands
- Elly Koeijers — Statistics Netherlands
- George van Leeuwen — Statistics Netherlands
- Frank van de Pol — Statistics Netherlands
- Robbert Renssen — Statistics Netherlands
- Jean Ritzen — Statistics Netherlands
- Christian Roques — Eurostat
- Peter Verboon — Statistics Netherlands
- Denis Ward — Australian Bureau of Statistics
- Winfried Ypma — Statistics Netherlands
- Kees Zeelenberg — Statistics Netherlands

The following persons at Eurostat have contributed significantly through their comments and suggestions:

- Simon Allen
- Raoul Depoutot
- Peter Struijs

Project co-ordination at Eurostat:
Jeroen Jutte and Christian Roques

For further information on this publication please contact Eurostat:

	Telephone	**Fax**	**e-mail**
Mr. J. Jutte	+352.4301 32032	+352.4301 32600	Jeroen.Jutte@eurostat.cec.be

CONTENTS

GENERAL OBSERVATIONS

I. AIM OF THE HANDBOOK

Statistical information is the final product of a more or less complex range of operations. In this respect statistics do not differ from more tangible products, like motor cars or dinners in a restaurant.

This handbook provides general guidelines on how to *make* statistics. It does not contain a comprehensive description of the *contents* of business statistics, nor is it a cookery book, providing an enumeration of recipes. Rather, it deals with cooking *principles* and *methods*, thus supporting survey statisticians in preparing their own recipes for making statistics, each from his own specific objective and infra structure.

The handbook applies for the setting up of new surveys as well as for the redesign of existing surveys. It is meant to serve several parties of interest. It supports survey managers of EU member states as well as transition countries that are moving towards the EU statistical system. And it provides training material for statistical courses, both at the EU and at the national level. Finally, it offers the professional users of business statistics a look behind the scenes, so as to enhance their understanding of what they read.

II. THE MISSION OF NATIONAL STATISTICAL INSTITUTES

Author: Ad Willeboordse

1. Recent developments

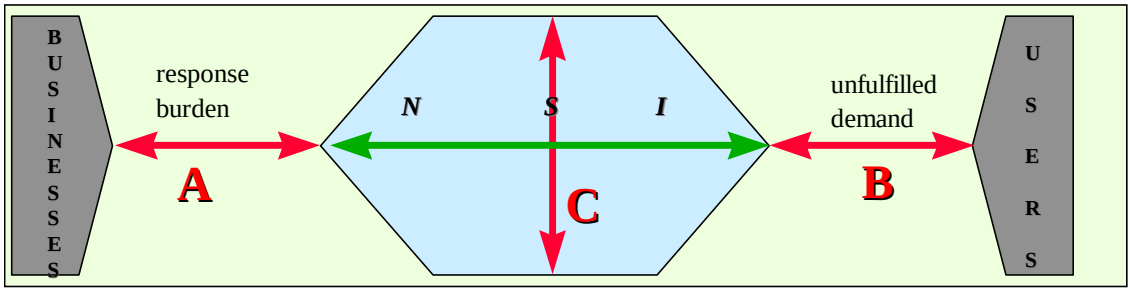
This Handbook is a guide for design *and* redesign of business surveys. While the first will in particular apply for transition countries, the latter is most relevant in countries with a tradition in statistics for a market economy. Presently, National Statistical Institutes (NSI) from all over the world encounter similar problems and opportunities, urging them to drastically reconsider their surveying strategies. Notably four developments appear to occur on a global scale:

- respondents become ever more reluctant to cooperate, while at the same time national governments announce programs to reduce paper burden on the business world;
- governments freeze or even cut budgets for NSI's;
- users become increasingly demanding, in particular with respect to timeliness, coherence and detail of information;
- there is a tendency towards growing competition on the information supply market; NSI's are no longer monopolists.

It is obvious that these four developments interfere closely, in the sense that the first two seem to hamper NSI's in adequately responding to the latter two.

2. NSI's tasks

The production of business statistics can be conceived as the bridging of the gap between information demand by users and information sources held by respondents. Ideally, an NSI should account for the whole bridging of the gap. In practice, it will never fully succeed. The traditional situation can be pictured as follows:



The figure shows that:

A. part of the bridging job is left to respondents, by asking them questions they cannot directly, i.e. without substantial effort, answer from their accounting systems. *Arrow A* represents the response burden, or, more positively stated, the contribution of respondents (in our case mostly businesses) to the statistical process;

B. the NSI does not *fully* meet user demand, as is shown by *arrow B*. Shortcomings in the statistical product may be of various nature: quantity, reliability, timeliness, consistency or accessibility of data may be less than required by users.

The horizontal arrow in the middle segment denotes the contribution of the NSI to the bridging of the gap between supply and demand. Arrows A and B can be said to denote (external) "tension indicators". The area of the "NSI-body" represents the funds available. Consequently, arrow C can be seen as a (negative) measure for the cost-effectiveness of the NSI contribution.

3. Challenges

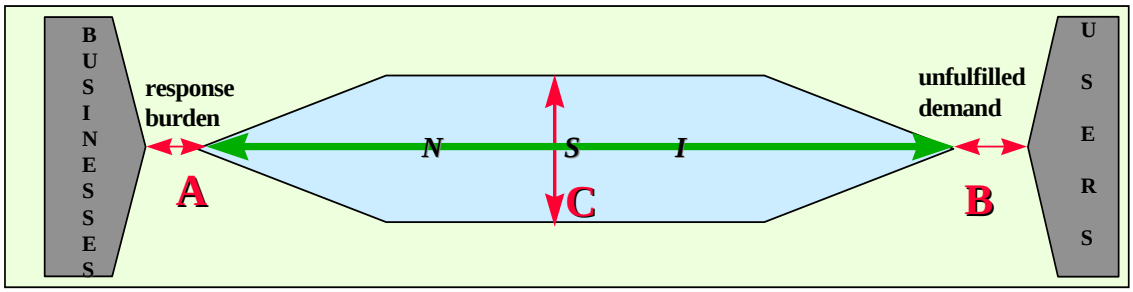
In terms of the picture, strategic programs should aim at a simultaneous reduction of the length of each of the three arrows:

A. response burden should decrease substantially, not only because governments require it, but also because this is the only way to ensure ongoing co-operation of the business world;

B. output value should increase, because users become more and more demanding and because of increasing market competition;

C. productivity should increase, because the funds with which the goals (A) and (B) have to be realised, will remain equal or even decline.

It is, therefore, the mission of the NSI to gradually change the picture towards the following form:



Notice the metamorphosis in the shape of the NSI-body: despite the loss of some weight (by ongoing budget cuts) it reaches further, both towards users and respondents. This is made possible by getting slimmer (more efficient) and consequently leaner.

Notice also that both users and respondents shift to the right; the first because of ever increasing demands, the latter because they are expected to be willing to adapt their accounting systems to a certain extent to statistical standards.

The picture accentuates the mutual dependency of the three performance indicators, which is indeed a crucial element in the whole process of transformation. Relief of the tension in one of the indicators should not lead to disproportionate increase of the tensions in any of the others. It would, e.g., be fairly easy to reduce response burden by doubling field staff and halving sample sizes. However, if the negative effects of such measures on productivity and output would exceed the positive effect on the burden indicator, such a transformation can hardly be qualified as a performance.

The question is, therefore, how to develop a strategy and to take measures which generate a positive effect on each of the three performance indicators, i.e. which both relieve response burden and NSI cost, and enhance the quality and the quantity of the output.

4. Strategies

In this Handbook we will pay special attention to those instruments and measures which are in particular fitted for this strategy. Among these instruments, the most outstanding are:

- the use of advanced information technology for data collection. Electronic Data Interchange (EDI) is expected to substantially reduce response burden and processing cost, while output will benefit, in particular with respect to timeliness and consistency;
- the use of administrative registers as alternative sources, also in cases where the data provided are not directly fitted for statistical needs;
- the use of advanced survey methodology with respect to editing, imputation and estimation, in such a way that processing cost fall and quality rises;
- the use of advanced information technology for data storage and dissemination, in such a way that users receive coherent information fitted to their needs, irrespective of the surveys the data are derived from.

Coherence of the statistical product is to be considered a key issue in this strategy, and all of the tools just mentioned support this goal. Users require data which can be related and compared with each other, whatever survey they originate from. NSI's can distinguish themselves from any other survey institute by the fact that they provide a comprehensive and coherent picture of the total economy, built up of numerous statistics, each of which is to be considered as a piece in the jig saw.

Users want to compare and relate data from different countries as well. This demands a coherent system which covers NSI's and provides standards for concepts and classifications. The European system of business statistics fulfils this role for a growing number of EU member states and transition countries.

III. REFERENCE TO THE EUROPEAN SYSTEM

Author: Christian Roques

1. Introduction

Although the Handbook focuses on *processes* rather than *products*, and therefore does not describe statistics as such, the guidelines are deliberately based on a conceptual and methodological framework. This framework is explicitly and implicitly embodied in a large number of EU-documents. Eurostat has developed a pattern of legal acts to ensure a rapid and reliable harmonisation of European statistics.

2. Status of Documents

It has to be noticed that these legal acts do not have necessarily equal status and do not create the same obligations towards Member States although some simple recommendations have a greater importance for the harmonisation process than Directives or Regulations .

The official EU ranking in terms of direct effect for a legal act is in a top-down approach the following:

- Council Regulation, Commission Regulation (serving as implementation of a Council Regulation)
- Directive of the Council
- Council Decision
- Recommendation

It has to be kept in mind that draft legal acts have no legal influence as long as they are not adopted and that certain Regulations are implemented thanks to explanatory notes, while manuals which do not have a mandatory legal status are merely advisory.

3. Relevant documents

This handbook is based on and supports the following legal acts:

Regulations:

- Council Regulation on structural business statistics
- Council Regulation on statistical units
- Council Regulation on business registers
- Council Regulation on labour costs
- Council Regulation on earnings
- Council Regulation on Intrastat
- NACE Rev.1
- CPA
- ESA

Directives

Directive on short term indicators for manufacturing and construction

Recommendations

- Recommendation on Small and Medium Enterprises

Manuals

- Manual of recommendations for drafting business registers
- Explanatory notes of NACE Rev.1
- Manual on short term indicators

Draft legal acts

- Draft Regulation on short term indicators
- Draft Commission Regulation on implementing the Council Regulation on structural business statistics.

In general, the above documents are referred to without repeating their contents. In a few cases, i.e. where there is a need for extensive comments or elaboration, major elements will be quoted.

IV. STRUCTURE AND CONTENTS OF THE HANDBOOK

1. Introduction

The Handbook follows as much as possible the subsequent stages to pass through when designing and implementing a survey, i.e. from monitoring of user needs to dissemination of the data. In practice, the sequence will be less straightforward than presented here. Indeed, most steps of the statistical process have an iterative character, so there will always be some redundancy, circularity and feedback among the different stages. One of the major causes for this is the need for a permanent balancing of user needs with survey cost and response burden. During all stages this trade-off will induce the survey designer to reconcile decisions taken in earlier stages.

The following stages will be passed through:

PART A

When considering a survey in a certain field of information, the very first activity to undertake is an *exploration* of demand and available supply in the area of interest. The result is a comprehensive inventory of user needs. Subsequently, these needs have to be translated into a specified device of the *intended statistical output.* This step is taken against the background of two very important benchmarks, i.e. the *general framework* of business statistics as embodied in the EU system and the *constraints* induced by limited *collectability* of the data desired. Complying with the first benchmark yields consistency and comparability of the data with those of other surveys. The second releases respondent burden and enhances co-operation and data quality.

The main features of the step from user needs to intended survey output are the delineation of the *target population* (including choice of statistical units and classifications) and the choice and definition of the *variables* about which data are to be produced. Final result of this stage is the *table outline*. This outline can be considered the representation and specification of the survey objectives, which is the eventual goal of this part of the Handbook.

PART B

This part deals with the *operationalization* of the survey objectives. Whereas the first stage was primarily *product* and thus user oriented, now the focus turns to the *process*, and thus to the NSI and the respondent. The question is how to achieve the survey goals with a minimum of cost for the NSI and burden for the respondent. Before proceeding to *survey* design, we will first discuss the situation where a survey can be avoided, i.e. the use of administrative registers.

Only after such registers have proven to be insufficient, the step towards the following preparatory operations applies:

- development of the *sampling design*, resulting in a sample selection scheme. This stage includes the choice of an appropriate sampling frame;
- design of the *questionnaire*.

PART C

This part discusses the actual implementation of the survey. Three operations apply: *drawing* the sample, followed by *collection* of the data and finally resulting in *entering* of the data in a micro data file. In view of the developments signalled in chapter II above, we will in particular focus on measures and policies for reduction of response burden and for increase of response rates and data quality.

PART D

The set of raw micro data is the starting point for the *processing* and *analysis* of the data collected. Consecutive steps are editing, imputation, estimation, integration and analysis. All these operations amount to storing of the data in a *data warehouse.* This data base comprises all statistical data, as defined in the table outline.

PART E

This data warehouse is actually the source from which the *tabulations* are derived. These can be considered the final products of the statistical process. A dedicated *dissemination policy* is necessary in order to make sure that the products reach a maximum number of users at the right moment in the right mode.

Before starting the cycle, we will provide a general review of the very first action to carry out, once the decision has been taken to consider a new survey or a redesign. This action comprises the planning and organisation of the (re)design project. Besides, we will discuss an aspect which leaves its mark upon every phase in the cycle and which is a major criterion in many decisions, i.e. the *quality* issue.

V. ORGANIZING AND MANAGING A SURVEY PROJECT

Author: Denis Ward

1. Introduction

The scope of all duties implied by a survey is so broad that deliberated project management methods are necessary to stage-manage the development of a new survey or a fundamental redesign of an existing survey. In fact, the activities applying during all segments of the cycle as presented in this Handbook have to be carefully arranged in advance. This chapter provides a brief review of organisational and management aspects.

2. Organisational aspects

From an organisational point of view, it is useful to divide the set of tasks to perform into two parts:

- those referring to *subject matter* issues, such as:
  - ⇒ statisticians in charge of surveys in the same economic *area* or on the same *theme* in other areas. These statisticians can be considered to represent the external users;
  - ⇒ internal "users". Here, one should not only think of National Accounts, but also of statisticians in charge of the policy for statistical co-ordination and for the dissemination of results. Besides, at the input side of the statistical process, contacts apply with centralised activities as the Statistical Business Register and the unit which is responsible for the response burden policy;
- those which are connected to the *equipment*, computer programs and staff needed to physically deal with the information. Choice of the mainframe/PC's and network technology, choice of adapted software, development of new programs and interfaces and training of clerical staff are some of the issues in this field.

In view of these different areas, one might consider to choose two heads of the project, i.e. the head of Statistical project and the Head of Computer Project. A Steering Committee can be raised, as well as one or more *project teams*, consisting of users, respondents, representatives of administrative registers and, of course, involved statisticians. Special procedures apply for contacting the Board of Directors of the NSI, in order to:

- secure efficient decisions in budgetary matters;
- inform about progress and delays;
- demand supplementary resources;
- ensure co-ordination with other departments.

3. Management techniques

The use of management techniques applies in order to ensure that the work is being done in a consistent and efficient way. It is important to take measures so as to limit the risks in the development. A normalised procedure enables to immediately replace the Head of the Project in case of illness, transfer, etc. Such a procedure also comprises the deadlines for the subsequent steps in the project. We will now list the main features of the management techniques:

Standardised steps

With respect to the management of the project, the following standardised steps can be distinguished:

- assessment of existing material;
- development of plans and sub-projects;
- realisation of the sub-projects;
- implementation of the new system;
- development of maintenance programs, documentation, archives.

Each step concludes with a report from the project managers, directed towards the Steering Committee. In case of agreement, the project managers are authorised to proceed to the next step. Moreover, at the end of the first and the second step, assessments of costs should be produced. The authorisation of the Board of Directors may be necessary at the end of these stages, due to the important budgetary aspects of these steps.

On-going planning and feedback

At each step, a comprehensive list of tasks to be performed should be defined. For each task a number of required characteristics is specified, such as resources for fulfilment, length and tasks to be finalised before. Subsequently, a complete diagram of each step can be produced, showing in a graph its duration and critical path. This continuous monitoring of time and resources enables to estimate the risk of delays in time tables and exceeds in budgets. On the basis of such information one can decide to make use of subcontracting so as to fight delays.

Assessment of risks

The reliability of information systems is usually a major concern. To avoid future problems in using or updating the system, risks have to be assessed at each step and proposals must be made to limit these risks.

Examples of frequently occurring risks are changes in software or replacement of hardware, and changes in the norms and standards of external data sources.

Tests

Every proposal and program should be tested, first within an experimental framework and subsequently within a real framework. Only after a successful test a task can be qualified as fulfilled.

Documentation

With information technologies advancing rapidly, it becomes more and more difficult for individual surveying statisticians to overlook and follow the whole statistical process. Therefore, the creation of an extensive documentation system is extremely important. Only with such a system the statistician can retain a proper understanding of the *meaning* of statistical concepts and of the statistical data. This documentation has to be built up during the stages of the design and implementation process. Frequent job transfers stress their necessity.

VI. QUALITY

Authors: Raoul Depoutot, René Huigen and Robbert Renssen

1. Introduction

In planning and design of business surveys, the issue of quality meets growing attention. Systematic approaches like TSD (Total Survey Design) and TQM (Total Quality Management) place the issue in the centre of survey design philosophies. Within the context of this Handbook, we have to confine ourselves to a brief review of the main features of the TQM approach. Besides, we will list and discuss a number of quality indicators. The quality measures, already requested by the European Commission in the framework of the EC Council regulation on Structural Business Statistics, are briefly presented.

2. The concept

The leading concept behind Total Survey Design is the *overall* optimisation of the statistical process as a whole. TQM can be considered an aspect of this concept, which states that all operations making part of that whole should be designed in such a way as to control all possible sources of error and to obtain the best possible precision in the survey results. The primary elements of TQM in the context of survey quality are therefore:

- the *collection* of information on key aspects of each stage of the design, collection, processing, and tabulation phases. Examples of such information include:

  ⇒ time taken by respondents to complete a questionnaire;

  ⇒ response rates;

  ⇒ error rates identified for each variable in a questionnaire;

  ⇒ lapsed time and resources required to undertake each of the survey phases, etc.;

- the use of this information in a systematic purposive way in order to obtain improvements in each of the elements measured. This process is generally iterative in the sense that large numbers of small individual opportunities for improvements are identified and implemented each time a (monthly/quarterly/annual) collection is undertaken;

- The involvement of staff actually undertaking the work. This is perhaps the most difficult aspect of TQM, as it may require a change in organisational culture to invest staff with sufficient skills and interest to use the information being collected to improve key processes. Indeed, the success and continuity of TQM is significantly influenced by change being driven by the people undertaking the work, rather than by management directive.

3. Aspects of quality

Each survey development process starts with explicitly stating the aims as formulated by the user groups and mutually arranged with the survey authorities. The aims include *relevance* of concepts, *reliability* of outcomes, processing *time* and survey *cost*. The best achievable survey is the one that satisfies all of these aims and makes optimal use of the available production means. In practice this quality target is never met, because the survey process is a human operation and errors are possible in each stage. The survey results realised include all imperfections of the process. Therefore, survey quality can be seen as the degree of approximation between the *best feasible* survey results and the results *effectively achieved.*

Aspects of overall quality can be derived from features of the survey undertaken, such as processing time, processing costs and statistical reliability of the outcomes. In general, time and costs can be easily measured. Statistical reliability, however, is often hard to quantify. The

statistical reliability of an outcome is determined by the total error. In case of a sample survey this error consists of a sampling error and a non-sampling error. Errors from different stages of the process cannot be seen independently from each other. An error made at the beginning of the process can affect subsequent stages.

To control survey quality, during any stage of the cycle information on the production process has to be gathered on a recurrent and systematic basis. To attain this information, additional provisions have to be made, such as special items in the questionnaire, the saving of subsequent versions of databases corresponding to different stages of the survey process (preliminary data, imputed data, edited data). In order to attain the best overall quality, all production means, such as organisation, management, employees, statistical instruments, computer facilities and training courses should be used as efficiently and effectively as possible.

The survey manager has to be informed on the survey quality during all stages of the process, in order to be able to adjust the process in time. The measured quality of the realised statistical outcomes can also be a signal to relocate the available production means and to obtain a more effective and efficient production process in subsequent surveys.

4. Quality measures

Quality is defined in the ISO 8402 - 1986 as: “the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs”. This could be analysed in the framework of Total Quality Management. We will adopt a more restricted approach in order to measure quality, concentrating initially on the “products” rather than on the process as a whole. Two exceptions will be made: relevance is a valuation and timeliness is related to data which are process information.

Quality of statistics can be defined with reference to several criteria:

Relevance

A survey is relevant if it meets users’ needs. The identification of users and their expectations is therefore necessary.

Accuracy of estimates

Accuracy is defined as the closeness between the estimated value and the (unknown) true population value. Assessing the accuracy of an estimate involves analysing the total error associated with the estimate.

Timeliness and punctuality in disseminating results

Most users want up-to-date figures which are published frequently and on time at pre-established dates.

Accessibility and clarity of the information

Statistical data have most value when they are easily accessible by users, are available in the forms users desire and are adequately documented. Assistance in using and interpreting the statistics should also be forthcoming from the providers.

Comparability of statistics

Statistics for a given characteristic have the greatest usefulness when they enable reliable comparisons of values taken by the characteristic across space and over time. The comparability

component stresses the comparison of same statistics between countries in order to evaluate the meaning of aggregated statistics at the European level.

Coherence

When originating from a single source, statistics are coherent in that elementary concepts can be combined reliably in more complex ways. When originating from different sources, and in particular from statistical surveys of different frequencies, statistics are coherent insofar as they are based on common definitions, classifications and methodological standards. The messages that statistics convey to users will then clearly relate to each other, or at least will not contradict each other. The coherence between statistics is orientated towards the comparison of different statistics which are generally produced in different way and for different primary uses.

About Cost

Although not a measure of quality, the resources available for the production of statistics act as a constraint on quality. When assessing the ability of a Member State to comply with quality guidelines, it is necessary to take account of the resources available. There is clearly a strong link between the quality of statistics and the resources available to produce them. An assessment of costs should be kept in mind during the quality evaluation process.

Accuracy

This aspect of quality has been extensively studied in many statistical agencies and by academics. Accuracy focuses on an analysis of errors, divided into sampling and non-sampling errors.

Errors can be classified as follows:
2.1. Sampling errors
2.2. Non-sampling errors
 2.2.1 Frame errors
 2.2.2 Measurement errors
 2.2.3 Processing errors
 2.2.4 Non-response errors
 2.2.5 Model assumption errors

Groves (1989), Lessler and Kalsbeek (1992), Särndal, Swensson and Wretman (1992), Biemer and Fesco (1995) use similar classifications.

Until now, statisticians have focused mainly on sampling errors. Estimates of accuracy published by statistical agencies usually cover only sampling errors (in fact the estimated variance depends in practice on response rates and certain non-sampling errors). However, research into non-sampling error is developing rapidly, and some methods are now available for a first assessment of such errors. For instance, see Biemer et al. (1991).

Timeliness and punctuality

Users generally require that statistical information takes a minimum amount of time to produce, is released as soon as it is available, and where appropriate is available on a regular basis.

Keeping production times to a minimum implies efficient production techniques, often at great expense. The whole process of data collection, editing, imputation, estimation and dissemination has to be kept under control in order to minimise the processing period. A partial solution to the problem is to provide early estimates, based on a sub-sample of respondents. Care needs to be taken over the presentation of these early estimates, and also with the subsequent dissemination of the revised information.

There is sometimes competition between (often private) statistical information, produced and disseminated very rapidly but with less reliability, and slower, but more accurate, statistics from government agencies. The trade-off between timeliness and accuracy has some very visible

consequences for certain users. This challenges statistical agencies to improve the timeliness of their statistics with no reduction in the high standards of quality.

Clarity and accessibility of statistics

Dissemination is a vital step in the information chain. It is not sufficient to have "good statistics" stored somewhere inside the statistical office. They have to be made available to all potential users, in an appropriate form. Firstly, users should be in a position to know easily which kind of statistics are available. Secondly, physical access to the statistics should be convenient. Thirdly, the statistics should be accompanied by the necessary information on concepts and methods. Different levels of explanation should sometimes be envisaged, in order to differentiate between those who are subject specialists and those who are not. Finally, analysis of the statistics may emphasise the service dimension of statistical products.

Comparability

There may be a difference between national concepts and European definitions. Any such differences should be described. For example, suppose that in a given Member State the concept of gross investment excludes the value of capital goods acquired through financial leasing but that they are included in the European definition. If it is not possible to collect at individual enterprise level the value of goods acquired through financial leasing, an estimate will need to be made at an aggregated level. Comparability is not limited to the comparability within the EU: Eurostat should be in a position to assess the comparability of EU statistics with the statistics of other countries or groups of countries (USA, Japan).

Comparability should also exist over time: changes due to modification of the reference concept or of the measurement process should be documented and the impact of these changes assessed. In the same way, changes in society in general (eg: new legislation, mergers and demergers) having an impact on continuity should be taking into account.

Coherence

Where similar statistics from various sources (surveys or statistics calculated from administrative data at the national level by the Official Statistical System) exist, they should be identified and any differences should, if possible, be quantified. A discrepancy between two sets of statistics produced by different surveys may be due to: differences in the data collection process or differences in reporting units resulting in different estimates. The situation may be improved by benchmarking (for instance monthly or quarterly statistics on annual results) or by combining different survey results. In any case, misunderstanding by users should be prevented by the use of different wording for the different concepts. In addition, statistics estimating complex concepts (ratios, elasticity, etc.) should be based on coherent elementary statistics (with compatible definitions of the reference population(s), characteristic, reference period, statistical unit).

Cost constraint

Quality is associated with costs. Information on the resources available to survey managers allows an assessment of how to optimise the budget they receive. The resources available to a Member State should be taken into account when assessing their likely ability to satisfy quality measures. Two components of cost can be considered*: the cost to the statistical office* and *the cost to the reporting units* (typically enterprises or parts of enterprises). The cost to respondents differs mainly according to the size of the enterprise and the sector of activity since different questions may be asked in different sectors. Costs for a new survey may be much higher than the costs for an existing survey. The use of administrative data drastically reduces the burden on enterprises.

Costs are more a constraint for quality improvement than a component of quality itself.

References:

Biemer, P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. and Sudman, S. (eds) (1991) *Measurement Errors in Surveys*, New York: Wiley

Biemer, P.P. and Fesco R. S. (1995) *Evaluating and controlling measurement error in business surveys*, in Cox and al. (eds) Business Survey Methods, New York: John Wiley, pp.257-281

Groves, R.M. (1989) *Survey Errors and Survey Costs*, New York: John Wiley

Lessler, J.T. and Kalsbeek, W. D. (1992) *Nonsampling Errors in Surveys*, New York: John Wiley

Särndal C.E., Swensson B. and Wretman J.(1992) *Model Assisted Survey Sampling*, New York: Springer-Verlag

PART A SETTING THE SURVEY OBJECTIVES

The first segment of the survey cycle comprises all activities directed towards defining the intended final product: i.e. a set of *tabulations* and supporting *meta data*, to appear in the eventual publication. The work starts with a systematic *exploration* and inventory of user needs. These are then *translated* into statistical concepts and visualized in tabulations. The process of translation is dominated by a confrontation of the user needs with two benchmarks: *degree of fit* with the standards of the general framework of business statistics and *feasibility* with respect to respondent load and processing cost. The first filter leads to modifications for the sake of coherence with data from other surveys and compliance with the European system. The second filter protects respondents from unbalanced overload and the NSI from setting unattainable targets.

I. THE EXPLORATORY STAGE

Author: Ad Willeboordse

1. Introduction

Before deciding on whether and how to conduct a survey, the statistician should have a clear view of both demand and available supply in the field of the information required. The first action to take is the identification of potential users in order to find out their needs and the uses they have in mind. Subsequently, one has to make sure that the information required is not already available from other sources. Thirdly, the potential data suppliers, i.e. entrepreneurs and their organisations, must be consulted in order to ensure that they have a positive attitude towards the intentions and to check their willingness to cooperate.

2. Identifying user groups

When the need for information in a particular field arises, it is important not only to communicate with the user (group) raising the request, but to involve *all* potential user categories which might be interested in the field. In general, the following main groups can be distinguished:

- Government.

  The information needed by government is diverse. It ranges from macro-economic information, needed for planning, economic and monetary policy, etc., to quite specific information. An important distinction applies for - mostly detailed - structural data versus quickly available and more aggregated short term indicators.

- Businesses.

  Operating in a primarily commercial context, businesses want to be informed on the behaviour of the industry group to which they belong as well as their suppliers and clients. Moreover, there will be a need for data on economic indicators such as inflation and economic growth.

- Research institutions.

- Statistical data are used in social sciences in general. Research institutes can operate on behalf of others, e.g. by analysing consequences of proposed policy measures, or can have a scientific aim. In general, the information needed by the government is also needed by research institutions, but the latter may require more detail.

- Associations, foundations, and other non-profit institutions.

- Like businesses, non-profit organisations need information about their environment. However, the aspects in which these organisations are interested generally differ from businesses. The emphasis is on social information related to the objectives of the organisation.

- Educational institutes.

  The needs of educational institutes are, of course, of a very diverse nature. Serving this group not only meets the demand for quantitative information on all aspects of society, but also educates pupils and students as future users by making them familiar with the existence and use of statistics.

- The general public.

  The needs of the general public are very diverse and concern all aspects of social life. The general public represents an important market for statistics and should not be overlooked.

- Foreign users.

  Apart from foreign individuals, businesses and institutions, there is a number of important international organisations which use statistical data for all kinds of policy purposes. Among the most important are the EU, UN, International Labour Organisation (ILO) and OECD. Separate mentioning of this category stresses the need for data to be internationally comparable. This asks, at a minimum, for harmonised concepts and classifications.

Last, but certainly not least, there are a few specific categories of interested parties, which can be labelled as *intermediate* users. We mention two of the most important:

- National Accounts.

  For most NSIs national accountants are internal users of business statistics. This category takes interest in data across the whole range of economic statistics. The development of an accurate set of national accounts is of paramount importance for government policy decision making, international comparisons, etc. One of the factors influencing the quality of national accounts is the accuracy and consistency of data from business surveys. It is therefore important for national accountants to be involved in the development of business statistics to ensure that the variables collected and classifications used are consistent with the European System of Accounts (ESA'95).

- EUROSTAT.

  The statistical office of the EU combines the member state statistics in European statistics and takes therefore a very large interest in harmonised data sources. For this reason, Eurostat is actually more than a client of NSI's; it is a partner and co-ordinating body as well.

Users can also be listed as *individuals*: statisticians, marketeers, politicians, managers, (market) researchers, students, journalists, teachers, datavendors, advisors, citizens. Such an identification can be useful when measuring satisfaction with existing statistical products.

3. Assessing user needs

Finding out what users really need is more than just letting them list a number of items or mark a number of pre-printed suggestions:

- the statistician should assist and guide the user to make up his mind and express his needs in unambiguous terms;
- to be able to do this the statistician must learn about the intended applications, in order to enhance her or his understanding of the meaning of the needs expressed and to judge their relevance;
- user needs tend to be infinite, in particular when these needs are measured without confrontation with their cost: everybody prefers more to less. Therefore, inventory of needs (by forms or interviews) without introducing any constraints should be avoided. By asking the users to rank their wishes in order of preference, a more realistic picture can be obtained. Furthermore, it is useful to distinguish between proven needs and potential needs;

- user needs on a specific topic should be considered within the context of the existing total framework of economic statistics. Users must make clear with which existing data they intend to relate or compare the data needed;
- often users will have conflicting needs: they want the data to be both detailed, accurate and quick. In that case they must be asked to make known their priorities;
- in case different user groups or users have related, but deviating needs on a certain issue, the statistician should try to establish consensus, instead of cumulating their wishes. Plenary consultations, during which users can exchange views, could be helpful. If choices have to be made, preference should be given to those needs fitting best in the general framework or harmonise with generally accepted standards, so as to ensure outcomes that are consistent with those of existing statistics. A second, but not less important measure is respondent-friendliness.

The last mentioned issue relates to a topical dilemma in modern statistical policies. On the one hand, we notice a growing tendency towards client-orientation, which actually asks for tailor-made statistical products. Whereas the latter may be well tuned to the specific wishes of a specific user, the "overall-client" who desires comparable data on all aspects of a certain theme, should not be overlooked. Thus, the need for tailor-made information has a counterforce in the need for standardisation, not only for the sake of harmonised statistics, but also to limit response burden. We will come back to this issue repeatedly. It will appear that modern information technology allows for modes of data capture and dissemination which serve both goals.

In a new field of information, user needs can be monitored by tracing and analysing client files of existing publications in related fields. User satisfaction can be established both by desk and field research. The latter can take the form of interviews, focus groups or questionnaires. Thus, more insight in additional needs is obtained. It is necessary to canalise the specific needs, because most of the data users have a broad interest. This can be done by letting the potential user specify the frequency (daily, monthly), the quantity (one figure, large number), the depth (micro data) and the purpose (to report, to teach) of data use.

Assessment of user needs should not be confined to statistical data as such, but include the mode of dissemination as well. In this respect, it is specially important to learn the preferences of users with respect to paper versus electronic media.

4. Exploring supply

Once a first indication of user needs being available, the statistician has to find out to what extent these needs could be satisfied by statistical information already available, either from resources within the own NSI, or from other data vendors. It is specially important to ensure that there will be no overlap with surveys conducted by other government agencies, research institutions, educational institutes, non-profit organisations, businesses (such as banks), and international organisations.

The notion of "data already available" should be taken in a broad sense. Existing data which do not exactly meet the needs, but on the basis of which the data needed can be calculated or estimated, should be taken into account as well.

After concluding that there is no or insufficient information available, the next step in exploring the supply market is to find out whether the data wanted can be derived from existing administrative sources, maintained for other than statistical purposes. VAT registrations are an important example. We will come back to this kind of alternatives later.

5. Contacting potential respondents

Only after obtaining full evidence that neither existing statistics nor administrative registrations are able to satisfy the needs, the NSI is entitled to contact individual enterprises and/or their trade organisations as potential data suppliers. A positive attitude of the potential respondents and their trade organisations is a major condition for a successful survey. To gain their willingness to cooperate, the following might be of help:

- entrepreneurs and organisations should be approached not only as respondents but also as users. If they are not interested spontaneously, their involvement might be awaked by giving them the opportunity to include their own particular wishes in the survey;
- often, the official publication envisaged is not specific enough to be of interest to the respondent. In that case, special issues should be released, preferably free of charge, fitting the needs of homogeneous (sub-)groups of respondents. E.g., data could be presented in such a way that one can easily compare its own position with the group where one belongs to;
- the NSI should, if at all possible, guarantee that the information to be collected will not go beyond data directly available from book keeping records, and that representatives of respondents will be narrowly involved when developing the questionnaire;
- the NSI should at any time be able to inform respondents on their *total* burden, as far as the own NSI-surveys are concerned. Often, respondents perception of their burden is higher than reality; such an overview *might* give some relief;
- the NSI should stress its confidentiality rules, guaranteeing that no individual data will be submitted.

6. Resource aspects

The inventory of both demand and supply enables the statistician to obtain a first indication of the total cost of a survey. When weighing this cost against the benefits, respondent burden should be taken into consideration as well.

Funds are not the only means to care about. Of course, one should make sure that human resources are available as well, both in quantitative and qualitative sense.

Finally, technical equipment, among which software, should be provided for.

7. Legal aspects

In case of a positive balance, and provided availability of funds, all conditions are met to put forward the survey plans to the competent committee. This committee, commonly composed of representatives of central government and all interested parties, will have to approve the survey and decide whether its status will become statutory or not.

8. Ongoing contact with users

An important element in ensuring that survey output meets the needs of users is the maintenance of ongoing contact with them throughout the survey development process. Users must be kept informed of progress and any changes in (intended) output that may arise during any stage of the design process (e.g. questionnaire testing), to check whether possible concessions are acceptable to them. Ongoing contact may be obtained through regular meetings with key users or through correspondence, inviting feedback to the NSI as survey development proceeds.

The need for such procedures underlines the iterative nature of the subsequent stages in the design process.

9. Summary

The exploratory stage comprises an exhaustive inventory of user groups as well as an investigation of user needs. These demands are matched with the possibilities of obtaining the data needed, either from administrative registrations or directly from enterprises. Provided the findings are positive, funds are available, and benefits are expected to exceed both internal and external costs, the first stage ends with the official consent by the competent committee.

II. THE FRAMEWORK OF BUSINESS STATISTICS

Author: Ad Willeboordse

1. Introduction

The exploratory stage ends up with a more or less detailed description of the information needed and the objectives the information is meant to serve. This description must now be translated into an appropriate statistical language that comprises concepts, variables, terminology and definitions. The result is a detailed specification of the intended statistical output.

Defining the statistical output for a particular survey is, however, not an isolated activity, exclusively based on an interpretation of user needs in the field of interest. We mentioned already that constraints with respect to data availability as well as considerations regarding response burden should be kept in mind. Moreover, in choosing the concepts, terminology, definitions, classifications etc., the desired or required relationship with other statistics must be taken into account. In this respect, the needs of National Accounts must be mentioned in particular. But there is more: in principle all statistics that allow for relations with each other must be considered. The whole of economic statistics can be viewed as constituting a comprehensive framework in which the concepts of the survey to design or redesign have to fit, at least as much as possible and desirable. Comparability and consistency are indeed principal features of the *quality* of statistical data. It goes without saying that comparability and consistency at the level of national statistics are not enough. Users require international harmonisation as well.

2. The need for coherence

User interests will generally not be confined to data resulting from one particular survey. Often users wish to compare or otherwise relate data from different survey sources. This tendency is enforced by the growing availability of on line data bases in which the data derived from a large number of surveys are stored and from which users can retrieve all data they wish, without being limited by the physical barriers of paper publications.

There are several kinds of obvious relationships to consider between data from different surveys:

- *additivity* of data referring to different sub-populations, e.g. employment in trade and in manufacturing industry etc. This is important for users interested in a certain aspect or theme of the economy (in this example the labour market);
- *comparability* of data on different aspects or themes for a certain area of the economy, e.g. employment data from labour surveys and turnover data from production statistics in the textile industry;
- *consistency* of the outcomes of short term statistics with those of annual statistics. Turnover data for a certain industry, measured on a monthly basis, should add up to turnover data observed by annual surveys;

- *consistency* of the outcomes of

  ⇒ regional statistics with those of national statistics, and subsequently

  ⇒ national statistics with those of supra-national statistics.

  Fixed capital formation data for the various regions of a country in a certain industry should add up to national data for this industry. These national data should in turn add up to aggregates on the EU level and UN and OECD level;
- *continuity* of the outcomes for period *t* with those for period *t+1*. The difference between turnover figures in Textile industry for December 1996 and January 1997 should reflect the *real* development of the industry;
- *fittedness* of concepts and outcomes of business statistics to the European System of Accounts (ESA'95) and the System of National Accounts (SNA).

3. Patterns of international harmonization

The ultimate goal is to achieve harmonised statistics throughout the world. Within this global perspective, the realisation of which is, of course, still far behind the horizon, different "levels" of harmonisation can be distinguished:

A. If the *European Statistical System* is the universe of discourse, two dimensions occur:

1. data supplied by EU member states should *add up* to EU-aggregates, according to the concepts of the EU- statistical system. We call this the *vertical* dimension of international harmonisation;
2. the various sub-systems within the EU-system should be mutually *consistent*, e.g.:

   ⇒ short term statistics with structural statistics;

   ⇒ industrial statistics with services statistics;

   ⇒ basic statistics with national accounts.

This we will refer to as the *horizontal* dimension of international harmonisation.

If we also take the *national statistical systems* of EU-member countries into consideration, there are two more relations to care for:

3. the concepts used in the publications of the national statistical institutes should correspond to those in EU-publications, so as not to confuse users;
4. national statistical institutes of EU-member countries should use identical terminology, concepts and definitions in their publications, thus enabling users to compare and relate data from different countries.

Finally, by expanding the universe of discourse to *global* harmonisation, two additional relation patterns have to be added:

5. The EU- statistical system should be coherent with the UN system;
6. national statistical institutes of *all* countries should use identical terminology, concepts and definitions in their publications, so as to enable users to compare and relate the data from different countries throughout the world.

Within the context of this Handbook we are primarily interested in the statistical framework as it is embodied in the European statistical system.

4. Conceptual basis of the framework

The relationships mentioned above can only be established in the presence of a coherent system of economic statistics, within which each survey is deemed to be subordinate to overall standards and concepts. As this Handbook focuses on practical issues, there is no room for detailed conceptual digressions. Therefore we will confine ourselves to a brief elaboration of the basics.

The basic idea behind the system approach is the notion of the economy as a circular flow of money and goods and services. Business statistics describe transactors and their transactions in the economy. These processes can, in principle, be observed in two ways, i.e. from the angle of the *transactors* in the economic process and from the angle of the *phenomena* to describe. We will call the first view the *actor* approach, as opposed to the *functional* approach to denote the second.

Example

The *actor* approach considers the computer services industry, as composed of the set of transactors (i.e. Enterprises or Kind of activity units) with principal activity NACE 72. This implies on the one hand that secondary activities outside NACE 72 are included, while on the other hand computer services, provided as a secondary activity by non-NACE 72 transactors are ignored.

The *functional* approach would focus on computer services, irrespective of the institutional or organisational environment within which these services are produced. Even *ancillary* computer services, rendered to parent or sister companies within the same business, will be involved.

In chapter III, where we discuss the delineation of populations, the position of the two approaches will be further explained. In this Handbook we will assume that the core of business statistics is organised and compiled according to the actor approach. In doing this, we follow the concepts of ESA, notably embodied in the *dual sectoring* and *dual actoring* concept. Indeed, one of the major implications of this approach is the central role of statistical units, representing the producing actors in the economy.

In most countries the framework is not explicitly described as a *system*. Rather it reveals itself as a set of generally accepted standards and methods, the use of which is recommended or compulsory. Whether or not a coherent system exists or is aimed at, in any case it is recommendable to try and comply with overall standards as much as possible when specifying the statistical output.

Powerful methodological instruments to attain coherence are:

- the use of standard statistical *units* as building blocks for aggregation of micro-data up to industries and other domains, like institutional sectors and geographic areas. Well known examples are the Enterprise Group, the Enterprise, the Local Unit, the Kind of Activity Unit (KAU) and the Local Kind of Activity Unit (LKAU). All of these units are listed and defined in the Community Regulation on statistical units;
- the use of standard *classifications* to classify units and to delineate populations, like NACE or ISIC for activities and the institutional sector classification of SNA for sectors. Regional and size classifications belong to the standards as well;

- the use of standard classifications for products, like the (EU) Classification of Products by Activity (CPA) and the (UN) Central Product Classification (CPC), the PRODCOM for production statistics and the Combined Nomenclature (CN) for foreign trade statistics;
- the use of uniformly and consistently defined core variables, relating to the principal transactions in the economic process.
- Statistical units, classifications and core variables can be considered the cornerstones of the framework, as well as instruments for co-ordination of the variety of economic statistics. Therefore, it is useful to briefly elaborate these concepts.

5. Sorts of Units

The notion of "units" applies in different stages and contexts of the statistical process. Discussions on the issue are often hampered by misunderstanding and confusion with respect to the "discourse" at hand as well as the terminology used. Therefore, it is important to clearly distinguish between those different discourses and to apply uniform terminology. Before dealing with the units as such we map the subject by presenting a typology of units. With respect to discourses, a distinction applies between:

1. The unit *from* which data are obtained and *by* which forms are completed. It is more a contact address or person than a real unit. This unit is often referred to as the REPORTING UNIT;
2. The unit *about* which data are reported. This is the unit to which the questionnaire, as completed by the respondent, refers. In general, it coincides with the unit which is at the basis of statistical aggregates, i.e. the building block of tabulations. Whenever EU Regulations mention units, they always mean this unit, and they use to call it OBSERVATION UNIT. Such units represent "real life" economic entities which can be identified, and about which data can be obtained. Enterprises, Local Units and Kind of Activity Units, as defined in the EU-regulation on Statistical Units, belong to this category;
3. Contrarily, there are artificially constructed parts of observation units, the so-called ANALYTICAL UNITS. They are "created" for the sake of enlarging (physical) homogeneity. The Unit of Homogeneous Production (UHP) is the most notorious example of this type.

Example

A computer retailer is drawn in the annual production survey. The owner leaves the completion of the questionnaire to his book keeping office.

1. The book keeping office is the REPORTING UNIT.

The computer retailer consists of three shops. The book keeping office completes questionnaires for each of the shops separately.

The annual production survey is, according to the CR SBS, based on Enterprises. This implies that the data reported at the level of the three shops, are consolidated (by the NSI) to Enterprise level.

2. The Enterprise is the OBSERVATION UNIT.

The retailer not only sells computers but also audio-visual products and magazines.

For National Accounts purposes (symmetric input/output tables) the Enterprise is splitted into three Units of Homogeneous production (UHP).

3. The UHP is the ANALYTICAL UNIT.

In practice the three unit types often coincide. Still, it is useful to distinguish between them, so as to avoid confusion and optical (dis-)agreements. Different contexts ask for different (types of) units.

In the context of this *conceptual* part of this Handbook we are primarily interested in observation units. Reporting units are relevant in the *processing* part, notably in the sampling and collection stage. Analytical units are beyond the scope of this Handbook, as they have little to so with surveys.

With respect to the terminology it has to be noticed that there is no world wide standardization as yet on the meaning of the terms introduced here. This goes in particular for the reporting unit. E.g., ISIC uses this name to indicate the unit that is called observation unit here.

6. Statistical units in the European system: concepts

Introduction

Statistical units play a prominent role in the European system of business statistics. Indeed, the CR on statistical units explicitly states that:

> "..... only if the member states use common definitions of statistical units will it be possible to provide integrated statistical information with the reliability, speed, flexibility and degree of detail required for the management of the internal market"

There is actually no Community regulation referring to business statistics which does not make mention of statistical units. This attention is fully justified, because statistical units are:

- the *corner stones* of the statistical system;
- the *building blocks* of statistical aggregates;
- the *linking pins* between the statistics to be harmonised.

The CR -SU deals with the conceptual aspects of statistical units, It lists and defines as much as *seven* statistical units and can be considered the methodological reservoir out of which the various regulations like ESA '95, Structural Business Statistics (SBS), Short Term Indicators (STI) and Statistical Business Registers (SBR) take the unit of their choice. Of the seven units mentioned, the observation units Enterprise and KAU, with their regional counterparts Local unit and Local KAU are the most relevant for business statistics: Notice that the *legal unit* is not listed as a statistical unit. Indeed, the legal unit is, in its own right, not relevant as an *output* unit in the system of economic statistic (which does not take away that it is an important *input* unit).

We will now provide a systematic review of statistical units, in such a way that the essential differences between the various types become apparent. Where the definitions of the units are not self-explaining, we will make use of the operationalizations as provided by the EUROSTAT document "The treatment of large and complex enterprises". This document serves as a guidance for the construction of statistical units in Business Registers.

The global level

The highest unit in the hierarchy is the Enterprise Group. Is defined in terms of (groups of) Enterprises..

ENTERPRISE GROUP

AN ASSOCIATION OF ENTERPRISES BOUND TOGETHER BY LEGAL AND/OR FINANCIAL LINKS

This unit is gaining importance with increasing globalization. Where all other units are confined to national borders, the Enterprise Group has no restrictions at all; it simply combines all legal units with common ownership and control. Indeed, LEGAL UNITS form its most elementary building blocks.

Example

Legal units a, b and c are located in countries A, B and C. Units b and c are owned (for more than 50 %) by unit a. This constitutes one Enterprise Group.

Notice that, logically spoken, a legal unit owned by "itself", constitutes an Enterprise Group too, however small and simple it may be.

The decision maker

As we saw, the Enterprise Group is made up of ENTERPRISES. However, the latter are bound to national borders. Still, their most important feature is that they represent more or less independently operating decision centres within the Enterprise Group. It is, therefore, very well possible that an Enterprise Group has several Enterprises within a certain country.

ENTERPRISE

THE SMALLEST COMBINATION OF LEGAL UNITS THAT IS AN ORGANISATIONAL UNIT PRODUCING GOODS OR SERVICES, WHICH BENEFITS FROM A CERTAIN DEGREE OF AUTONOMY IN DECISION MAKING, ESPECIALLY FOR THE ALLOCATION OF ITS CURRENT RESOURCES.

The definition does not contain any reference to homogeneity according to activity or location. On the contrary, the explanatory notes deliberately state that:

"an Enterprise carries out one or more activities at one or more locations".

The Enterprise corresponds to the *institutional unit* as mentioned in ESA'95.

For a proper understanding of the Enterprise concept it is also important to keep in mind that it *can* consist of several legal units. Combining legal units which belong to one ownership applies when only the combination meets the requirement of "autonomy in decision making".

The question remains when to combine LUs in one Enterprise and how far to go. This question has been answered by EUROSTAT in the earlier mentioned document "The treatment of large businesses". We will follow this document by illustrating three possible situations:

1. LUs within an Enterprise Group, providing internal ancillary services to other parts of the Enterprise Group;
2. LUs organised according to vertical integration;
3. LUs organised according to horizontal integration.

1. Example ancillary units

Assume an Enterprise Group, originally consisting of one legal unit and engaged in the production of motor cars. The business takes care for the transport of the cars by itself.

Assume further that the transport activities are put in a separate legal unit, which only serves the parent unit and which has no autonomy of decision making.

The two legal units together make one Enterprise.

However: If the transport unit would service more than one Enterprise within the group, it was to be considered a separate Enterprise, although not being an independent decision maker. There are two reasons for this deviation from the definition: a Legal Unit may never comprise more than one Enterprise and respondents may have difficulties to allocate their services to each of the ENT's served.

2. Example vertical integration

An Enterprise Group has a printing division put and a bookbinding division, both put in separate LUs. All printing work is binded by the bookbinding division; all books binded are printed by the printing division.

The two Legal Units together make one Enterprise.

The two processes are not independent; only when combined they constitute one autonomous decision making unit. The fact that the printing unit, whose complete output is consumed by the bookbinding unit, is not market oriented, is an important indication for non-autonomy.

3. Example horizontal integration

Suppose an Enterprise Group made up of three legal units: one for the production of wooden tables, one for wooden chairs and one for cabinets.
The operations are technically and economically strongly interwoven: there is one management, the resources are shared, inputs are combined and marketing is done for the business as a whole.

The two Legal Units together make one Enterprise.

It would be highly artificial to consider them as independent decision centres, even if data were available at the level of LUs.

Notice that, in the last example, the "wholeness" of the decision making transactor is considered important enough to resist the temptation of splitting for the sake of homogeneity (remember that the NACE distinguishes classes for chairs and other furniture). This "holistic view" is fully justified, not only because the Enterprise definition demands it, but also because it is unlikely that *meaningful* and *realistic* data at the level of the three legal units would be available.

More homogeneous, but still real life

The KAU is meant to reduce the heterogeneity according to activity which is inherent to the Enterprise. At the same time, it avoids devaluating to an artificial construct.

KIND OF ACTIVITY UNIT (KAU)

THE KAU GROUPS ALL THE PARTS OF AN ENTERPRISE CONTRIBUTING TO THE PERFORMANCE OF AN ACTIVITY AT CLASS LEVEL (4 DIGITS) OF THE NACE

AND CORRESPONDS TO ONE OR MORE SUBDIVISIONS OF THE ENTERPRISE.

THE ENTS' INFORMATION SYSTEM MUST BE CAPABLE OF INDICATING OR CALCULATING FOR EACH KAU AT LEAST THE VALUE OF PRODUCTION, INTERMEDIATE CONSUMPTION, MANPOWER COSTS, OPERATING SURPLUS AND EMPLOYMENT AND GROSS FIXED CAPITAL FORMATION.

For a proper understanding of the KAU concept it is important to realise that the second part of the definition together overrule the first part of the first sentence: secondary activities are indeed allowed. Taking this into account, one might argue that the *naming* of the KAU is somewhat misleading. The KAU is indeed a true *observation* unit for which, *by definition,* data on operation surplus and other production items are available. It is actually the transactor in the production process[1].

[1] The KAU might fairly well correspond to the *business unit*, a term from modern business parlance

Example

1. Suppose an Enterprise that produces coffee and milk. In the sense of the NACE, these are two different activities. It is clear that the production processes are very different. Consequently, it is likely that these processes are managed and marketed independently from each other, and that there is a need for separate accounts. In other words, distinguishing two Cause would be a reflection of *economic reality*.
2. Suppose an Enterprise that produces wooden tables and chairs These are also two activities in the sense of NACE. However, there is a fair chance for separate data on operation surplus etc. not being available from the company's accounts. Production processes of tables and chairs are likely to be interwoven in many respects. Splitting would be artificial, as would be the (undoubtedly painfully) collected data. In such cases, we would conclude to one KAU, classified according to its principal activity, while taking the secondary activities for granted.

Notice that a KAU *can* consist of more than one legal unit. This follows logically from the fact that the KAU is derived from the Enterprise without the legal structure being a discriminating criterion. The fact that a certain Enterprise may consist of more than one legal unit, can therefore never be sufficient justification for splitting the Enterprise into two or more KAUs.

The EUROSTAT document on the treatment of large Enterprises recommends to consider splitting only when each of the following conditions is met:

In case of vertical integration:

1. the integrated activity is not listed as a single NACE activity at the 4 digit class level;
2. the activities of the legal units themselves are listed in NACE at the 4 digit level;
3. accounts are available with input and output data at (approximate) market values.

In case of horizontal integration:

1. the legal units carry out different economic activities in terms of NACE at the 4 digit level;
2. accounts are available.

The Community Regulation on Statistical Units explicitly states that ancillary activities will by no means constitute KAUs.

The ultimate homogeneity

Not being hindered by data availability constraints, the Unit of Homogeneous Production (UHP) is perfectly homogeneous in terms of CPA-groupings, i.e. NACE 4 digit classes. It is, therefore, not an observation unit: the data have to be produced by manipulation by the statistician. Its definition justly indicates that the UHP is more an *activity* than a *unit*:

UNIT OF HOMOGENEOUS PRODUCTION (UHP)

A SINGLE ACTIVITY WHICH IS IDENTIFIED BY ITS HOMOGENEOUS INPUTS, PRODUCTION PROCESS AND OUTPUT.

The UHP is artificial; it does not refer to any physical, accounting or economic reality. Therefore, its significance relates more to variables to collect than units to use in the observation process.

The regional dimension

Except for the Enterprise Group, all units discussed so far were delineated according to national borders. These units do not apply for sub-national statistics. Therefore, the Enterprise, KAU and UHP need a regional counterpart: Local unit, Local KAU (LKAU) and Local UHP (LUHP).

LOCAL UNIT

AN **ENT** OR PART THEREOF, SITUATED IN A GEOGRAPHICALLY IDENTIFIED PLACE.

- Like the other regional units, the Local Unit is a physically delimited unit; it is therefore independent of the regional classification applied;
- the Local Unit can consist of more than one Legal Unit;
- the Local Unit can comprise more than one activity, though in practice it will often be more homogeneous than the Enterprise.
- though a derivate from the Enterprise, the Local Unit does not inherit its main characteristics: it is not an autonomous decision making unit, and consequently data availability is limited.

LOCAL KIND OF ACTIVITY UNIT (LKAU)

THE PART OF A KAU WHICH CORRESPONDS TO A LOCAL UNIT

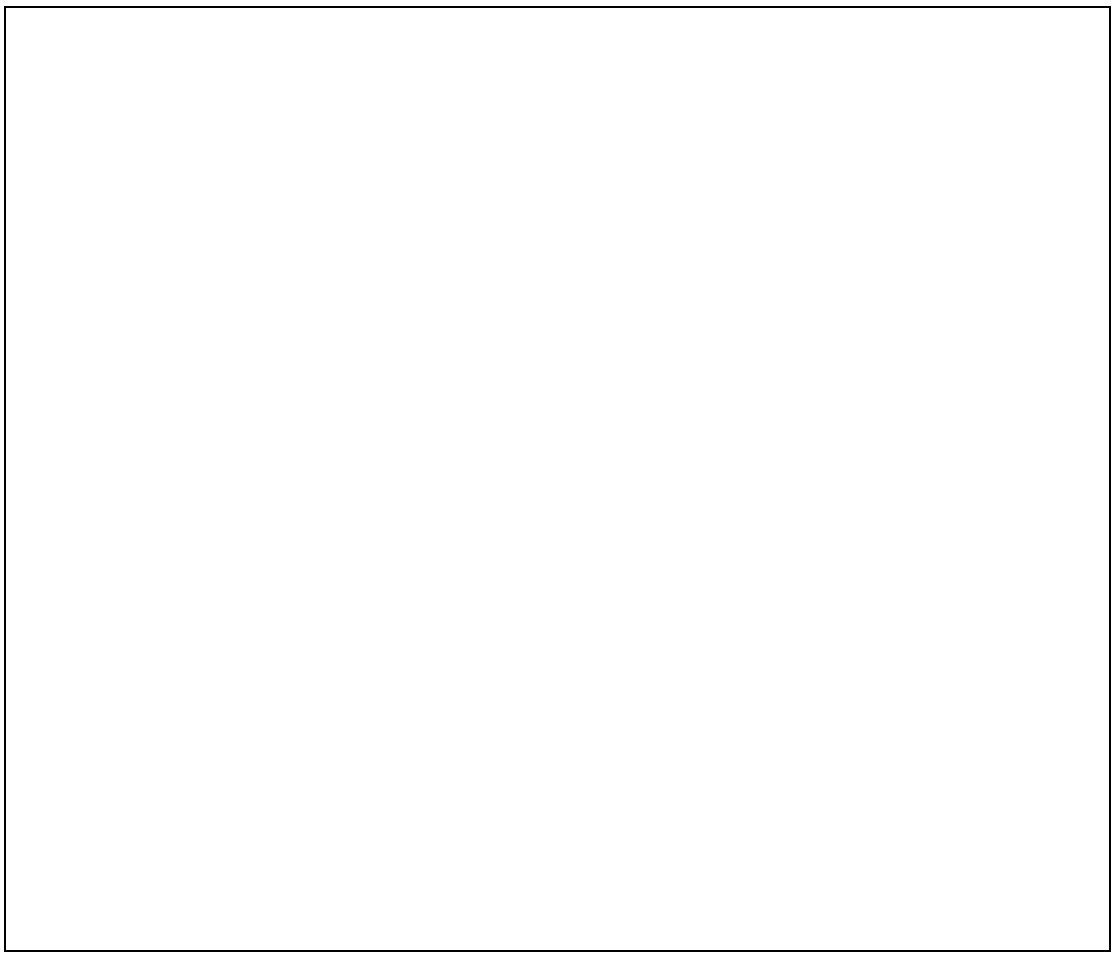
The LKAU can be looked upon both as a part of a KAU and as a part of an Local Unit. It is true that the LKAU is an *observation* unit; however, data availability is poor when compared with the KAU. Often, only data on employment and gross fixed capital formation can be obtained.

LOCAL UNIT OF HOMOGENEOUS PRODUCTION (UHP)

THE LUHP IS THE PART OF AN UHP WHICH CORRESPONDS TO A LOCAL UNIT

The figure below illustrates the unit structure of a particular Enterprise Group, *constructed out of:*

- 4 LUs, which are made up of 5 affiliates (local parts of Legal Unit) and subsequently *divided into*:
- 3 Enterprises, 4 KAUs, 4 LCU’s and 5 LKAUs.



Balancing cost and benefit

The fact that the statistical units Enterprise and KAU do not follow directly, i.e. in a 1 : 1 relationship, from the legal unit, has considerable consequences for the maintenance of Statistical Business Registers. Therefore, decisions to implement conversion activities from legal units to Enterprises and subsequently to KAUs should only be taken after careful analysis of the benefits and the cost. EUROSTAT recommends to confine n : 1 conversion from Legal Unit into Enterprise to EGs with 250 or more employees. Of course, such a recommendation does not take away that the operational rules as explained above fully apply, i.e. are in line with the CR SU.

7. Statistical units in the European system: uses

The statistical units discussed above provide the building blocks for the aggregates of business statistics. Therefore, each of the Regulations dealing with these statistics makes reference to one or more of the units dealt with. The following overview shows which regulations make use of which units.

Community Regulation	Statistical unit applying for:		
	financial data	production data	
		national level	regional level
Structural business statistics	**Enterprise**	**Enterprise**	open
Short term indicators	n.a.	**KAU**	n.a.
Labour cost statistics	n.a.	**Local Unit**	**Local Unit**
Structural labour statistics	n.a.	**Local Unit**	**Local Unit**
ESA'95 financial accounts	**Enterprise**	n.a.	n.a.
supply and use tables	n.a.	**LKAU**	**LKAU**
symmetric i/o tables	n.a.	**UHP**	**LUHP**

From this overview it appears that the various regulations use different units. This is not only the case for different sub-systems within the total system, but also within sub-systems. E.g., for production data at the national level four unit-types occur. It is in particular the use of the Enterprise and its local counterpart Local Unit, which might at first sight create some surprise. Indeed, when it comes to aggregation of production data according to NACE categories, the KAU is superior to the Enterprise for logical reasons. While the explanatory notes in the Regulation on Statistical Units concede that there may be some secondary activity within it, it can *by definition* be expected to be more homogeneous than the Enterprise to which it belongs. It is also true that statistics compiled on the basis of KAUs will be less affected by changes in the activity pattern within an Enterprise. This means that the aggregates compiled on a KAU basis will permit more valid comparisons of the activity associated with a process both between countries and over time.

In practice, however, it may take substantial cost and effort to identify the KAU in Business Registers. For this reason the Regulation on Statistical Business Registers requires the maintenance of Enterprise and Local Unit in stead of the KAU and the LKAU. Still, it stands to reason to aim at replacing the Enterprise/Local Unit by the KAU/LKAU in those cases where it really matters, i.e. in the case of large multi-activity enterprises. Of course, this goes for production data only; the Enterprise remains the appropriate unit for observation of financial data and aggregation to institutional sectors, as is explicitly stated in ESA'95.

8. Standard classifications

Overview of classifications

While statistical units are the building blocks for statistical aggregates, classifications are the adhesive. They provide the criteria according to which the building blocks are united into populations. For business statistics, the most commonly used classifications are:

- ***activity classifications.***

At the European level, the EU NACE REV. 93 is the standard, while at the world level, the UN ISIC applies. The former is an extension of the latter, so that the two standards are compatible. This implies that (European and non-European) countries using NACE, implicitly use ISIC.

The KAU and the LKAU are the logical observation units for aggregation of production data according to NACE categories. Of course, other units like the Enterprise and the Local Unit can be classified according to activity as well. However, this will lead to less homogeneous classes.

- ***institutional sector classifications.***

These are specially important for National Accounts purposes, in order to classify financial data for the Financial Accounts. It is, indeed, SNA/ESA'95 which provides the standard. Here the appropriate building block is the Enterprise, being the unit corresponding to the (ESA) institutional unit.

- ***product classifications.***

These differ fundamentally from activity and sector classifications in that their object of classification is not a statistical unit but, as the term says, a product. Unlike activities, the products are defined irrespective of the way their production is organised in economic entities. Two of the most important general product classifications are THE CENTRAL PRODUCT CLASSIFICATION (CPC) of UN and the CLASSIFICATION OF PRODUCTS BY ACTIVITY (CPA) of the EU. The EU PRODCOM is not a classification but rather a *list* of products to be used in production statistics.

The CPA groups the products according to NACE classes. Still, it should not be conceived as a mere extension of NACE (in which case its value as a *product* classification would be limited). Indeed, a particular product cannot be attributed intrinsically to one and only one NACE category.

- ***size classifications***

Size is an attribute of statistical units. It is mostly expressed in number of persons employed. The object of classification can be any statistical unit. Like the previously mentioned classifications, there is a European standard, be it that it is more or less hidden on a place where one would not expect it: the CR on SBR comprises a size classification to which is referred in the CR SBS.

- ***regional classifications*[2]**

At the EU level, the absolute reference is the classification of territorial statistical units (NUTS). Regional classifications are important for numerous industrial sectors that encounter different business cycles according to the territorial unit envisaged. Regional classifications are as well of importance for the elaboration of regional accounts. Since 1988 NUTS is used into the EC legislation as a regional benchmark, for instance in the Council Regulation related to structural funds.

NUTS is composed of 3 levels most often strongly linked to the administrative reality of Member States. Nevertheless, as most of the Member States only have two major territorial

[2] Author: Christian Roques

levels (Länder and Kreise in Germany, standard regions and counties in the UK), and as these levels are rarely comparable in terms of size and economic importance among Member States, a third level was introduced in the NUTS. This leads to a certain extent to the creation in each Member State of one artificial or at least unusual territorial level, since one level at least is not a major administrative territory.

NUTS is regularly revised to take into account the enlargments of the European Union and keeps thanks to several adaptations a strong economic validity.

- ***classification of changes***

A rather special classification refers to *changes* which units - and business populations - can undergo. Major events are birth, death, change of activity or size, and structural changes like merger and split off. Such a *classification of changes* is useful for tabulation in economic demographic statistics and for analysis of data based on subsequent populations that undergo substantial changes in their composition. Showing the effects of such changes enhances the understanding of time series, by isolating the effects of population changes.

Example

When total employment in computer services industry increases by, say, 10%, the understanding of this figure can be considerably enhanced by providing information on the effects of changes in the population. This answers questions like:

- what was the share of really new born firms in this increase?
- what was the influence of "new" enterprises which previously existed as computer service departments of industrial enterprises (management buy-out)?

This subject has been dealt with intensively in the context of CR SBR, notably by the development of a recommendation manual. In this Handbook, the treatment in Business registers of changes in units and their attributes is dealt with against the background of a classification of *events*.

9. Populations of units

Statistical aggregates for business statistics use to relate to groupings ((sub-) populations) of objects which are delineated according to one or more of the classifications mentioned above. The objects are either statistical units or products. There are notably three types of groupings which are often confused with one another, i.e. sectors, industries and branches. Therefore, it is important to clearly define and distinguish these concepts.

INSTITUTIONAL SECTOR

An institutional sector is a set of institutional units (*enterprises)* with similar financial behaviour, grouped according to the (SNA) institutional sector classification. Thus, the notion of institutional sector primarily applies in the discourse of financial data and financial accounts.

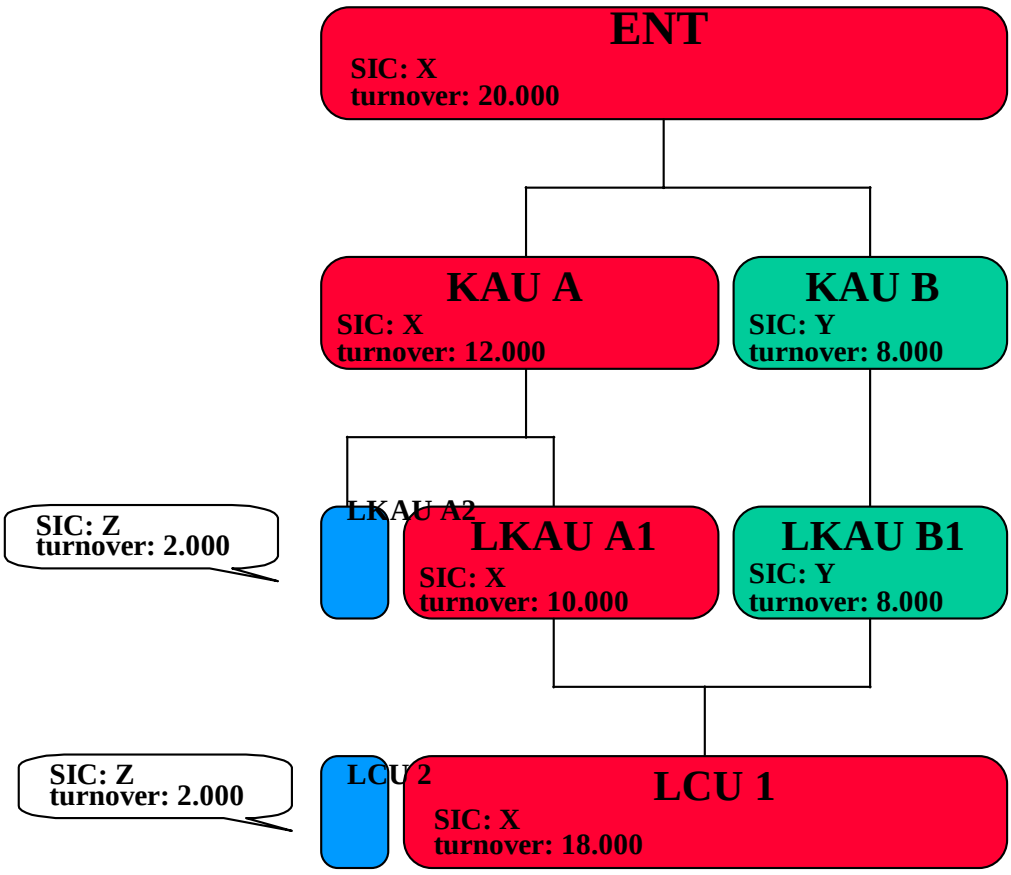
Example

An Enterprise Group carrying out both banking and business services will be split up into two Enterprises, one belonging to the sector financial enterprises and the other to the sector non-financial enterprises.

INDUSTRY

Obviously, NACE Rev.1 provides the classification instrument to delineate industries. However, neither the NACE Regulation nor the SU Regulation prescribes which statistical unit is deemed to be the *object* to be classified. In fact, the various EU-regulations on statistical output base their NACE-aggregates on different units, which means that the concept of an *industry* has various contents.

The example below illustrates how different units generate different "industries" and result in different data.



The question is: what is the contribution of Enterprise E to industry X? The answer depends on the choice of the unit:

Statistical unit	Turnover in Industry X
Enterprise	20 000
KAU	12 000
LKAU	10 000
Local Unit	18 000

We noticed already that neither the CR -NACE, nor the CR -SU comprise explicit statements with respect to the unit to be preferred as the basis for the INDUSTRY concept. However, ESA '95 mentions - in the context of make and use tables - an INDUSTRY as being composed of LKAU*s:*

"an INDUSTRY *consists of a group of LKAUs engaged on the same, or similar, kind of activity"*.

It is, however, not possible to give the LKAU the status of a *standard* unit, because of its limited data availability. Indeed, neither of the REGS-SBS, STI, LCS mentions the LKAU as the basis for statistical aggregates.

A remedy could be to classify the LKAUs according to the SIC code of the mother-KAU. In our example LKAU A2 would be reclassified from Z to X. The result would be that KAU and LKAU based NACE aggregates correspond to each other (both show 12.000). In that case, discrepancies are confined to KAU-Enterprise-Local Unit based figures. In any case NSI's have to adopt a policy that prevents users from being confused by different industry concepts and consequently different data.

(HOMOGENEOUS) BRANCH

A branch is a set of UHP's producing similar products (as identified by CPA), where the criterion for similarity is that the products are characteristic for one class of NACE at the 4 digit level. Notice the fundamental difference with the industry. Where the first is composed of real-life observation units (actors in the economy), the second is an artificial construct, the use of which is primarily found in symmetric input-output tables of National Accounts. To avoid confusion, it is strongly recommended not to mix up the terms branch and industry; they belong to different worlds indeed.

10. Key variables

Although business statistics cover a broad and diversified area of issues, still it can be said that there are a limited number of key variables that form the core of the system. The following scheme shows these concepts and their mutual relations.

	VARIABLE	SMALLEST UNIT APPLYING
	SALES / TURNOVER	KAU, LKAU, break-down by product
minus	trade inputs	
	PRODUCTION VALUE	KAU, break-down by product
minus	intermediate goods and services	KAU, limited break-down by product
	GROSS VALUED ADDED	KAU
minus	personnel cost	KAU, LKAU
	indirect taxes	KAU, LKAU
	indirect subsidies	
	GROSS OPERATING SURPLUS	KAU
minus	capital cost / depreciation	KAU, LKAU
	(fixed capital formation)	KAU, LKAU
	NET OPERATING SURPLUS	KAU
minus	interest	
	rent	
	other income / revenues	
	PROFIT / LOSS	Enterprise.GROUP/Enterprise

These concepts, one or more of which occur in a variety of surveys, should be labelled and defined in a uniform and consistent way throughout the range of business statistics.

Example

Comparability requires that, e.g., the concepts "number of employees" and "personnel cost" handle the item "temporary employees" equally in their respective definitions; it should not at the same time be included in one and excluded from the other variable.

Another type of comparability refers to the use of sub-concepts: in the survey on R&D the item "number of R&D staff" should be derived from the mother-concept of "employees".

The maintenance of a central list of standardised variables and definitions can be a useful instrument for co-ordination in this respect. Such a list does not only enhance coherence among different statistics, but also guarantees uniform language and unequivocal meaning of terms as used in publications. This can be obtained by stating that variables mentioned on the central list may only be used in the meaning as expressed by their official definition.

The use of most of the key variables mentioned above is made compulsory via the EU regulations on Structural Business Statistics and Short Term Indicators. Thus, these regulations act as supra-national rulers of standardised concepts, the compilation of which is compulsory for EU countries. The "Methodological Manual of Structural Business Statistics" and the "Glossary

of Business Statistics" and "Methodological Manual of Business Statistics" contain listings of variables.

11. Variables and statistical units

So far we discussed units and variables in separate paragraphs. However, from the foregoing it will be clear that the selection of variables and the choice of statistical units are interdependent actions, and therefore should evolve simultaneously. In general it can be said that the lower the unit in the hierarchy of which the Enterprise Group is the top level, the less data can be obtained or meaning-fully compiled. The relationship between units and variables can thus be characterised as a trade-off between homogeneity on the one hand and number and detail of variables on the other. The right column in the scheme above shows the "smallest" units applying for the various core variables. E.g., the variable "profit" can be measured at the level of the enterprise or enterprise group only; there will be no meaningful profit and loss accounts available at lower levels. This implies that statistical data on profits, observed at ENT level and tabulated according to NACE-categories, will inevitably suffer from a certain degree of "pollution".

Data on value added can be measured at the level of KAUs, and consequently bear more homogeneity. Core variables that can be meaningfully measured at the level of regional units, like the Local Unit and the LKAU are confined to employment and fixed capital formation mainly. One has to be careful with compiling regional statistics on, say, turnover. Even when respondents would be able to supply such data from their record keeping systems, there might be reasons to ignore these. Internal flows between production units and their own - physically separated - sales units should not be included in turnover data. This would lead to double counting and unrealistic data in case the deliveries are not charged at market value. Indeed, the requirement of "data availability", which plays such a crucial role in the definition of the KAU, refers to *relevant* and *meaningful* data only.

As pointed out before, the most homogeneous unit, i.e. the (L)UHP, is defined and used for analytical purposes and is of an artificial nature, since no requirement of observability applies. Data for such units are created in the statisticians' laboratories, in particular by National Accountants.

12. Summary

The value of statistical data will increase when outcomes can be compared with data from other surveys. Therefore, in specifying the statistical output, one should take into account existing standards and methods, whether or not these are subject to regulations or embedded in a coherent system. Complying with these standards not only enhances comparability, but suppresses uncontrolled growth of vocabulary as well, while at the same time reducing costs and efforts of survey development. The CR SBS and the CR STI contain a comprehensive listing of standard key variables, and thus provide a substantial contribution to the national and international comparability of business statistics. Moreover, the concepts are defined in such a way that the data resulting are useful for National Accounts as well.

With the general framework in mind, we will now turn back to the discussion of the survey design process, i.e. the stage comprising the translation of user needs into a specification of the intended statistical output.

References

Struijs, P., and Willeboordse, A.J., 1995, Changes in Populations of Statistical Units, In: Cox, B. et al, Business Survey Methods, Wiley , New York, pp 65-85.

III. SPECIFYING THE INTENDED STATISTICAL OUTPUT: THE TARGET POPULATION

Author: Ad Willeboordse

1. Introduction

The inventory of user needs, placed in the context of the standards and norms of the general framework, provides a solid and sound basis for the definition of the intended statistical output. Taking into account that a statistical figure commonly relates to a variable (e.g. turnover) and a domain of interest (e.g. the computer services industry), our discussions will concentrate on these two major issues. In this chapter we focus on the definition of the target population as the physical representation of the domain of interest. The next chapter will deal with the selection and definition of variables

2. Levels of populations

The target population for a certain survey represents the domain of interest for that survey. The appropriate tools for delineating this domain are the statistical units and the classifications dealt with in the previous chapter. Together with the time dimension they enable to specify the population envisaged.

Example

Market research by the NSI of country X has revealed that users need structural data on the computer services industry on a yearly basis. They want the data to be as homogeneous and comprehensive as possible and therefore they want them to be based on a homogeneous statistical unit and they want them to cover the informal economy as well.

The example describes the ideal target population which fully meets user demand. However, there are reasons and circumstances which make it necessary to deviate from this ideal population, which to a certain extent may even be qualified as *utopical*. Thus, there are actually *two* target populations to consider:

A) the *ideal* target population, which is a perfect representation of the user wishes;

B) the *intended* target population, specifying what users can expect to be delivered. When we speak of the target population in this Handbook, we mean the intended one.

To complete the picture, a third and even a fourth population level have to be added:

C) the frame population, i.e. the set of units actually involved in the survey and from which the sample is taken;

D) the sample, i.e. the set of units from which data are actually collected.

It is important to have a clear notion of the differences between the four layers. These will be discussed now by following the route from a) to d).

A→ B FROM IDEAL TO TARGET POPULATION

There are essentially three reasons to deviate from user wishes:

- overruling of these wishes by the standards of the general framework;
- benefits do not balance with cost and efforts, including those of respondents;
- adaptation for "impossible" needs.

Example (continued)

- For their specific purposes, certain users may dislike the criteria and rules for classifying computer services businesses as embodied in NACE. However, satisfying such a demand would disturb the coherence of the system as a whole, as well as comparability with data for other industries and for other countries.
- Although the need for comprehensive data is recognised, it may be too difficult or costly to collect or even estimate reliable data for the informal businesses.
- Although users may wish data on operation surplus at the most homogeneous level, i.e. the items of the CPA (which indeed many users do), it is considered not justified to do so, not even by calculation or estimation.
- Although KAUs obviously would be the second best alternative, it may be considered too burdensome to base the survey on KAUs, so that actually Enterprises are chosen.

Taking these considerations into account, the NSI decides to define the target population for the reporting year 1996 as:

the set of all Enterprises officially registered at any moment in 1996, as far as they are active and classified with principal activity "production of computer services", according to NACE Rev.1.

It is, of course, not possible to provide general guidelines with respect to the question how far to go with deviating from user wishes. This must be judged for each specific situation. However, in all cases it is very important to be *aware* of the deviations and to properly *document* them. In certain cases, it is also necessary to explicitly mention deviations in the publication. In our example, when presenting data on the computer services industry, one should inform the user that informal businesses, although belonging to this industry, are excluded (of course, the better thing to do is provide an estimation).

B → C: FROM TARGET POPULATION TO FRAME POPULATION

Ideally, the frame population, which is mostly derived from a Business Register, corresponds to the target population. In practice, it will often differ, simply because no Business Register is perfect. There will always be a certain degree of both overcoverage and undercoverage.

Example (continued)

1. Undercoverage

The Business Register of country X may lack recently updated lists of non-employed enterprises. This undercoverage must be compensated by *estimation*, based on other information sources, e.g. extrapolation of censuses.

2. Overcoverage

The Business Register may comprise a considerable amount of "enterprises" which actually died before 1996. Correction for this overcoverage may be undertaken by means of survey findings.

C → D: FROM FRAME POPULATION TO SAMPLE

In the majority of business surveys, only part of the frame population is represented in the actual data collection. A sample is drawn in such a way that the population data can be estimated by weighting the sample results. There may be situations where certain groups of units are excluded from sampling.

Example (continued)

Although enterprises which died somewhere in 1996, belong to the frame population, the NSI decides to exclude them from sampling because it is unlikely that their data can be obtained by collection. The data for this group will have to be estimated separately.

Frame and sample populations are being dealt with in more detail in PART B. In the context of PART A, we are primarily interested in (ideal) target and (intended) target populations (A and B).

To summarise:

- The ideal target population perfectly meets user needs.
- The target population represents the actual objectives of the survey (the intended statistical output). In presenting the statistical outcomes, no attempt is made to bridge the gap with the ideal target population.
- The frame population is the set of statistical units, providing the best possible materialisation of the target population. The gap between the two populations is bridged by alternative sources and calculations.
- The sample population is the set of statistical units, actually involved in data collection. The "gap" with the frame population is bridged by weighting and estimation.

3. Sub-domains of populations

The highest level of a population for national business surveys is the set of all "businesses" which are active at the national territory at a certain moment or during a certain period. As we saw in the previous paragraph, users are not only interested in data at the level of the *total* population, but require break downs as well.

Delineation according to kind of activity

Obviously, the most widely used viewpoint for breakdown is the *economic activity* (according to NACE or a national extension of NACE). As pointed out before, the KAU is the "natural" observation unit for groupings according to NACE (which we call industries), as this statistical unit represents the transactor in the production process *and* as it is the most homogeneous unit for which key data on the production process can be obtained. However, we noted also that there may be good reasons to yet apply another unit, notably the Enterprise. There may certainly be cases where reasons of cost-effectiveness justify the choice of the less homogeneous Enterprise. Still, it is important to have a clear understanding of the impact the use of different units may have on statistical outcomes (as we have shown in Chapter A-II-9).

If **ENTERPRISES** are chosen to be the elements of the target population, the computer services industry consists of all enterprises with principal activity computer service industry. As the Enterprise unit is not necessarily homogeneous with respect to activity, all non-computer service industry activities of computer service enterprises are included in the population (and consequently in the data observed), while (secondary) computer service activities of (primary) non-computer service enterprises will be disregarded.

When, for whatever reasons, the Enterprise is chosen as the statistical unit, it is recommended to at least partly compensate for the distorting effects of the "pollution" by providing for a detailed output specification according to products.

The **LEGAL UNIT** - though not a true statistical unit, nevertheless often used as such for the lack of better - shows in principle the same pattern. Only when multi-activity enterprises would incline to split up their different activities into separate legal units, the effects described will be more moderate. Special attention should be paid to the widespread habit of businesses to put up their computer department in a separate legal unit. If such a legal unit only renders services to sister or parent units, its activity is essentially *ancillary*. In legal unit based surveys one might wish to exclude such units from the population of the computer service industry, because they are not considered transactors on the market, i.e. their services are not rendered to third parties. This can be done by assigning them the SIC code of the unit(s) they serve.

If the computer services industry is considered to consist of **KAUs,** the computer services activities of non-computer services industry enterprises are part of the computer service industry, but - as the KAU definition states - only as far as these are independently operated and meaningful data on operating surplus and its components as well as capital investment data are available. Indeed, as stressed earlier, also KAUs can carry out secondary activities, be it to a lesser extent than ENTs. KAU-based populations will show more homogeneity and a higher coverage ratio with respect to the NACE category applying than Enterprise or legal unit based populations.

Although a regional unit by nature, the **LKAU** is a candidate as well (as we noticed before, according to ESA it is the appropriate unit for production data, both at the regional and the national level). Compared to the KAU, the use of the LKAU *might* increase homogeneity, i.e. in so far KAUs carry out their secondary activities on different locations. This also goes for activities of an ancillary nature: the computer department of a manufacturing Enterprise or KAU, which is separately located (in another street, city or region), will be considered an LKAU. For reasons to be explained later, it is recommended to classify such an ancillary unit both according to its own activity and to the activity of the KAU it belongs to.

The last EU unit to mention is the (**L**)**UHP**. It differs from the KAU, in that secondary activities always generate separate units. As a consequence, the number of units in the population will be considerably higher than for any other unit mentioned so far, while homogeneity and coverage ratio's would, in principle, rise up to 100%. The UHP as a unit is closely related to the product classification.

The ultimate degree of physical homogeneity is attained when one also wishes to consider *ancillary* activities as elements of the population. This would mean that all computer departments or even just staff involved in computer services make part of the computer services "industry". This is an example of a purely *functional* approach of business statistics, as opposed to the more common actor approach.

Delineation according to geographic area

Next to activity, a frequently applying viewpoint for breakdown is the *geographic area*. In surveys where kind of activity is not relevant, the sub-population for a specific region is defined as the set of Local Unit active in that region. In most cases, however, the regional breakdown is *additional* to the activity breakdown. Then the LKAU is the preferred unit, provided practical constraints.

The regional units Local Unit, LKAU and LUHP relate to each other in the same way as their national counterparts Enterprise, KAU and UHP relate to each other. So in fact the comments of the previous sub-paragraph apply as well, in the understanding that the relationship between regional populations, based on regional units and national populations, based on national units, demands special regard. It is important to notice that the first do not add up to the latter.

Example (continued)

When a KAU with principal activity manufacturing and secondary activity computer services industry falls apart into two LKAUs, one of which carries out the manufacturing and the other computer services, the latter LKAU is represented in one of the regional LKAU-based populations, but not in the KAU-based national populations. This implies that the sum of regional, LKAU-based employment data differs from national, KAU based data.

In principle, there are two ways to restore consistency:

- One could base both regional and national data on the same unit, i.e. the LKAU. In this case the problem remains that other variables, like operating surplus, cannot be observed meaningfully at the level of LKAU, so that co-ordination at the national level between these data and employment data would be distorted.
- One can also accept the dualism of the two units, and assign all LKAU the SIC code of the KAU they belong to. This might, however, be difficult to accept by those users which are interested in regional data only.

Therefore, what is the best solution depends on user demand. It might well be that there is a need for both approaches: present co-ordinated and "uncoordinated" tabulations as well.

It should be noted that the first approach never applies in case of ancillary regional units. These should always be classified according to the KAU(s) they work for.

Other criteria for delineation

While activity and region are the most frequently used criteria for delineation, additional criteria are often applied to further limit the size of the intended target population. E.g., the *size*

classification is often used to exclude smaller businesses, mostly for reasons of economy. Exclusion of smaller units from the *target* population implies that no estimation is made with respect to their share. In that case users should be carefully informed about the existence of this gap. It is, of course, in all cases better to try and make a calculation of the missing part. Although smaller size classes usually account for only a small share of total performance of industries, they can be important nevertheless, e.g. for early monitoring of turning points in the business cycle and as labour market indicators. .

Another limiting criteria is the *institutional sector*; this classification is often employed to exclude non-market units from the population.

4. The time dimension

Periods and moments

Next to classifications and units there is also the time reference to delineate a population. In principle, populations refer to either a *moment* or a *period*. Moment-bound populations apply well for the compilation of moment-bound data like employment or stocks. However, many variables, like turnover and fixed capital formation, have the character of a flow and consequently refer to a period. This raises the question how to deal with population changes in the course of such a period.

Example

A computer service firm is born in February 1996. After considerable initial capital investments the firm is not successful and dies in November 1996. Although the unit was absent in the populations of both January 1 and December 31, it has to be fully accounted for in the capital investment survey for 1996 (and other surveys).

Therefore, the appropriate population of the computer services industry in 1996 can be defined as the union of *all* populations of any moment in 1996. So, turnover of computer services industry in 1996 comprises the data of all KAUs or Enterprises, existing *at any moment* in 1996, with major activity NACE 72.

With respect to populations of the period-type a number of variations occur, ranging from a week to one or even more years. There are, in principle, no linking problems between data referring to different periods: turnover data from the 12 monthly surveys of 1994 are consistent with the results of the annual survey (provided no other errors). There is, however, one condition: the populations should be strictly defined as mentioned above. This goes specially for units that died during the reference year; they should be included in the population for the annual survey.

A special case occurs when a survey contains both moment-bound and period-bound variables. Now two different populations apply, the moment-bound one being a subset of the period-bound population. As the composition of the two populations is different, the two types of data cannot simply be related.

Example

Calculating the turnover (period-bound) per employee (moment-bound) requires some kind of manipulation of the employment figure, e.g. averaging with the figure for the previous year. Another solution would be to confine oneself to calculation of the ratio for the subset of units, making part of both populations. In case of seasonal patterns it is recommended to base the average on data for each of the four quarters.

Longitudinal populations

This type of populations makes a special use of the time dimension, i.e. by adding a *historic* property as a criterion for selection.

Example

A population may be defined as "all KAUs classified as computer service industry *both* on Jan. 1, 1996 and Jan. 1, 1997".

Such populations apply specially for certain forms of time series analysis, for which one wishes to eliminate the effects of changes in the composition of the population. Here, the categories of the earlier mentioned *classification of changes* can be of use.

Cohorts like *enterprise panels* are outstanding examples of longitudinal populations, their composition being frozen during a number of years. The criteria for selection of such panels can be manifold. They are not necessarily confined to "static" aspects like activity, region or size. Also, for certain types of analysis, dynamic aspects apply.

Example

An example is the tracking over time of "all KAU, born in 1995 with principal activity computer service industry". Such an exercise might apply in studies where the proposition is checked whether new and small firms generate more employment than older and larger ones.

5. Supra-national populations

For an NSI the total population of "businesses" is delineated according to national borders. As we have seen, businesses can be considered at different levels, like Enterprises and KAUs: One should, however, be aware that national borders are more and more artificial criteria in defining units. This goes in general as a consequence of globalization and in particular for EU countries, within which economic frontiers have recently faded away.

One of the consequences is the growing importance of the ENTERPRISE GROUP, which is the only one of the seven units of the CR - that is not restricted by national borders. Still, as long as *national* statistics are to be compiled, *national* units are to be used. It is, therefore, recommendable to pay special attention to the data supplied by businesses, whose financial or production activities are strongly interwoven with their foreign mothers, sisters or daughters. Special arrangements with respondents apply with respect to the reporting of transactions with these "third parties", so as to avoid that non-realistic allocations between parts of such conglomerates appear in statistical data.

6. Summary

Classifications and statistical units are the main tools with which the appropriate target population can be delineated. Choices should be consciously made, because they can have a considerable effect on the figures as well as on homogeneity of data. The time dimension must not be overlooked when specifying the population, in particular when one wishes to eliminate the effects on time series of certain changes in the composition of the population.

In making choices with respect to populations, co-ordination aspects should be taken into account as well, in particular relating to consistency between regional and national data.

IV. SPECIFYING THE INTENDED OUTPUT: THE VARIABLES

Author: Ad Willeboordse

1. Introduction

As pointed out earlier, the delineation of the target population (including sub-populations) and the selection of the variables to observe are closely related operations in the process of survey design. The statistical unit can be considered the linking pin: which unit or units apply depends on both the viewpoints for delineation and classification envisaged and the type of variables to observe. Therefore, the reader should be aware that the actions described here actually occur simultaneously with those described in the previous sections of this chapter.

The output specification should always precede the questionnaire design; the questionnaire results from the table outline and not the other way around. Despite noble intentions, it is true that it requires a fair amount of discipline to hold this rule in practice. Still, we stress on it without any reservation. Holding this rule does, however, not mean that one has to stick to the *original* output specifications throughout the design process. We mentioned already that the design process has in fact an iterative character. Pilot testing of questionnaires, discussions with potential respondents etc. may very well lead to modification of the intended output as originally specified. Failure to do so beforehand may easily result in an inability to translate data items collected in the questionnaire into a set of tables that meet user demands.

Next to selection of variables, we will discuss various ways to define variables, as well as considerations to take into account when choosing an appropriate terminology.

2. Selection of concepts

Starting point for selection of the variables to be observed is, of course, the information need resulting from the exploratory stage. These needs should first be matched to the concepts of the framework of economic statistics, as described in CHAPTER II of this part. National accounts concept play an important role in this respect. To avoid unnecessary diversity, one should try to conform as much as possible to the standardised terms and definitions of the "central list of co-ordinated concepts and definitions". Deviations may lead to needless divergence from other statistics, in which case comparability will be reduced and thus the value of the data.

Conformation to standardised concepts and definitions also relates to concepts which are derivates from other (mother) concepts.

Example

If fixed capital formation is defined as a standardised concept, the survey on capital formation in environment should start from this definition.

3. Definitions

As contents are more essential than names, the making of definitions should, in principle, precede the choice of vocabulary. This implies that user needs relating to certain concepts, should first be expressed in terms of definitions, after which the appropriate labels are assigned.

There are different ways to define concepts. C*onceptual* and *intentional* definitions refer to the essence, the purpose or the function of a concept:

Example

- "The KAU is the actual transactor in the production process";
- "Fixed capital formation is the acquisition of durable production goods.

One should avoid using circular definitions, merely "translating" a term in synonyms, thus failing to add real information.

Example

- "Turnover is sales to third parties".

Such a definition is only useful if the terms "sales" and "third parties" are further elaborated.

Conceptual and intentional definitions mainly apply for communication with users, both in the stage of inventory of user needs and in publications. Such definitions are an expression of the relevance of the concepts in relation to the uses envisaged, and as such they are the starting point for the derivation of operational and extensional definitions. These are necessary to attain a clear and common understanding of the precise meaning of a concept.

An *operational* definition specifies a concept, so that it can be applied in concrete situations:

Example

- The KAU is a part of an Enterprise for which data are available on operation surplus, etc.
- "durable production goods are goods to be used in the production process during more than one year".

Such a denotation is not only unambiguous, it also provides easy decisions rules, releasing statisticians from fundamental and tiresome discussions on the conceptual level.

An *extensional* definition comprises an exhaustive listing of all the elements, belonging to the domain of the concept. Such definitions, which mostly take the form of explanatory notes and which are often supplemented with exclusions, are specially useful for communication with respondents. Taking this purpose into account, it stands to reason that the listings should as much as possible be expressed in the form of book keeping items to be included, as well as items to be excluded. This implies that the design of an extensional definition should start with an inventory of *possible* elements (e.g. wage components, as distinguished in wage accounting systems), after which for each of the elements it is decided whether it should be included in or excluded from the definition. The conceptual definition is the benchmark in this process.

When developing extensional and operational definitions for use in questionnaires, it is important to be aware that different groups of respondents (e.g. small as opposed to large businesses or manufacturing as opposed to service industry) have their own *couleur locale* and therefore require their own operationalizations and vocabulary.

Defining concepts in terms of their basic elements has two major advantages: they leave little room for diverging interpretations and they enable the statistician to show differences and correspondences with related concepts, which are defined in terms of partly the same and partly other components. The use of book keeping items as components has one more advantage: it forces the statistician to take into account the reporting constraints of respondents. The latter does not mean that *output* concepts should always be defined such that they can be built up exactly from book keeping items. One can decide to deviate and obtain the missing items by imputation, either to be undertaken by the respondent or by the statistician (which of the two applies, is decided on in the questionnaire design stage).

If it is considered not justified or worthwhile to collect or even impute the missing item, it can be excluded as yet from the output concept. Such departures from the "ideal concept" should, however, be described in the explanatory notes of the publication or as a footnote to the table.

4. The European system

The METHODOLOGICAL MANUAL OF STRUCTURAL BUSINESS STATISTICS provides an ANNEX in which a number of core variables are defined provisionnally, both conceptual/intentional and operational/-extensional. The definitions inform the reader about:

- the meaning of the concept;
- the contents of the concept (by inclusions and exclusions);
- the link to national accounts;
- the link to other variables;
- the link to business accounts.

The GLOSSARY OF BUSINESS STATISTICS lists and defines in alphabetical order more than 200 variables in the field of business statistics. Next to variables, a number of key words of the statistical systems are mentioned.

5. Terminology

When assigning the most appropriate label to a given definition, *uniformity* of (statistical) language to apply towards users is the first thing to care about. One should avoid to employ terms already used with another definition in other statistics. In such situations one must either conform to the other definition or look for another term. For reasons of standardisation, the first solution is to be preferred, provided it is acceptable in view of the objectives of the survey.

Accessibility of statistical presentations can be enhanced by using terms which are familiar in common parlance. Such terms are, however, often vague and ambiguous as to their precise meaning. Concepts like “number of employees” or “enterprise” have numerous denotations. The use of such terms is, of course, only justified when the definition envisaged lies within the boundaries of usual interpretations of the term. Sometimes a definition is so specific, that it would be misleading to use a common parlance term. Then it makes sense to create specific vocabulary. This is especially true when the meaning of a concept is crucial for the understanding of the data. An example is “LKAU” (in stead of establishment).

6. The role of reporting constraints

Like with defining the target population, also with respect to variables the statistician has to take into account feasibility. Here, the willingness and capability of the respondent is an important constraint to take into account. This does, however, not mean that the intended output should in all respects *conform* to respondents willingness and capability. The latter is rather a condition for the *questionnaire*, being the eventual communication medium between the NSI and the respondent. It is, therefore, very well possible that the concepts, terms and definitions applying for the communication with users, differ from the concepts, terms and definitions as specified in the questionnaire. It is the task of the statistician to bridge the gap between data collected and data required. Therefore, the statisticians’ task to take into account respondents can, in the stage of defining the output, be understood as: do not set the level of ambition of the intended output so high that the gap between output variables and accounting systems of respondents is too broad to be bridged during the processing stage.

7. Summary

The choice of the variables for which data are to be published, as well as their definitions and terminology, is, of course, in the first place a matter of user needs. However, there are more aspects to take into account. The choices made should as much as possible comply with the features of the general framework of business statistics, as embodied in, i.a., a central list of concepts. Moreover, reporting constraints should be taken into account, be it to a certain extent.

Conceptual definitions are necessary but not enough. They must be accompanied by operational and extensional definitions, so as to be unambiguous and to leave no room for varying interpretations. Applying a uniform terminology towards users is a corporate issue, touching on the whole of business surveys.

V. DESIGN OF THE TABLE OUTLINE

Author: Ad Willeboordse

1. Introduction

All activities and choices described in the previous paragraphs of this chapter eventually culminate in the design of the table outline. This is in fact the visualisation of the features of the final product: a well organised set of clearly defined concepts, classified according to various viewpoints and expressed in a number of tabulations. The table outline is for a statistician what the building scheme is for an architect. It provides the benchmark for the forthcoming stages in the survey design. And it supplies future users with a clear picture of the type, detail, timeliness and accuracy of the information the survey is intended to supply.

2. Contents

By defining the reference period, by delineating the population and by selecting the variables, the boundaries of the survey scope have been drawn. Table design asks for decisions with respect to criteria for break down of the data and the level of detail to be presented to users. The most obvious criteria for tabulation are the *standard classifications* of the general framework. The classifications most widely used refer to economic activity (NACE), geographic area and size. It is important to specify the desired level of detail already in the design stage. The same goes for the choice of the *aggregation functions* according to which micro data are transformed into statistical data: counts, sums, averages, ratio's etc. Finally, the *dimensions* into which the data are to be expressed have to be determined: levels, time series, values, quantities, index numbers, etc.

The well known classifications mentioned above are not the only viewpoints according to which data in tabulations can be broken down. Provided their availability in business registers, also the economic demographic attributes, as distinguished in the classification of changes, can be used. An example is a break down of surveyed data according to age of the units. Furthermore, it is possible to use the variables measured in the survey as a viewpoint for classifying the population. E.g., splitting up the population - and the data - in profit-makers and loss-makers considerably enhances the information value of data on the economic performance of industries. Another example relates to a breakdown into classes of turnover, rather than employment.

3. Meta data

The worth of statistical data not only depends on their relevance as such, but also on their reliability and the time it takes to process and issue the data.

Standards as to timeliness and reliability should not be the unplanned residue of a survey, but a precondition, set during the design stage. A challenging goal is that the time of delay of the first issue should not exceed the length of the reference period. For example, turnover data referring to the month of January should be issued at March 1 at the latest. Provisional data are an approved means to relax the tension between the conflicting demands of timeliness and reliability.

Standards with respect to timeliness and accuracy should be part of the decisions concerning contents and scope of the survey. In particular the conflicting demands of timeliness and detail on the one hand and accuracy on the other must be balanced in advance, preferably by consulting the users.

The table outline should be accompanied by a detailed description of ambitions with respect to reliability and timeliness. When provisional data will be issued, the outcomes for which this is

intended should be specified. Also, for each statistical outcome a quality indicator of the total error must be provided, when feasible.

Such meta information not only intends to inform the user, but also specifies the level of ambition to be taken into account in the survey design, the planning and the implementation stage.

4. Initial and eventual table outline

We stressed already several times that the survey design process is to a certain extent of an iterative nature. This is specially true for the table outline. It is very well possible that the eventual outline will differ under the influence of the steps to be described in the following chapters, including the data collection phase. Therefore, users should be warned explicitly that the final tables may differ from the initial ones.

Example

The level of classificatory detail users eventually receive may differ from the initial outline, because it is simply not always possible to precisely predict the distribution of responses across all cells in a table. One approach may be to specify detailed initial tables that will be collapsed when survey results are tabulated. Such collapsing might be necessary for reasons of confidentiality or unacceptably high standard errors.

5. Summary

The systematic overview of user needs, resulting from the exploratory stage, provided the starting point for a detailed specification of the intended statistical output. The general framework for business statistics backs the statistician with a set of standards and methods. The choices eventually made with respect to the delineation of the target population and the choice and definition of variables eventually condense in the tabulation outline. This specification of the final product is not only defined in terms of its contents, but of its quality as well.

Table outline and quality standards are the lead for the next stage in the design process, of which sampling and questionnaire design are the major features.

PART B PREPARING THE SURVEY OPERATIONS

The previous stage amounted to a detailed specification of the intended statistical *product*. In the stages to follow the emphasis will be on the statistical process, i.e. the *input* and *throughput*. Part B focuses on two preparatory activities, related to the data input: design of the sampling strategy and design of the questionnaire. These two design activities are in fact the operationalizations of the two major features of the statistical output, i.e. the target population and the variables to observe.

In an ideal world without any constraints as to budgets, a world with a perfect business register and model respondents who are fully willing and able to report on all data desired, this part of the manual would have been brief and simple. However, in practice none of these conditions is ever met, so that a further *operationalization* and *transformation* of output specifications applies:

- business registers will not contain a perfect representation of the target population, so that special attention of the survey designer is required. Appropriate measures have to be taken in order to cope with the shortcomings of the business register;
- in most situations it is neither possible nor justified to carry out a comprehensive census, not only for reasons of budget constraints, but also for the sake of timeliness and respondent burden. This asks for a sampling strategy, aiming at a maximum reliability of outcomes at a minimum sample size;
- whereas output concepts and definitions are primarily *user* oriented, the questionnaire is the medium to communicate with respondents, and should therefore be fully *respondent* oriented. This implies that output concepts have to be translated into terms and definitions complying with accounting and management information systems. As to those items for which this is not possible, the statistician should deliberately look for alternatives like imputation or calculation, before deciding to load respondents with troublesome questions and burden himself with doubtful answers.

The scenery set above is based on the assumption that information is satisfied by means of a (new) survey. This is, of course, only true after a thorough investigation of existing administrative sources has proven that these do not to provide a suitable alternative. Therefore, we start PART B with a discussion on the pro's and con's of the use of administrative registers.

I. USE OF ADMINISTRATIVE REGISTERS

Author: Peter Bøegh-Nielsen

1. Introduction

The importance of administrative registers for use in the compilation of statistics is rapidly increasing under the pressure of three general developments, i.e. the need to reduce response burden and the budget cuts imposed by national governments on the one hand, and increased demand on the other. The role of such registers in the (re)design of business surveys can be briefly summarized by stating that, given a particular user need, direct data collection from businesses is only justified after ample investigation has led to the conclusion that - in terms of cost-benefit optimalization - none of the registers available can satisfy the need.

2. The essentials of administrative registers

In the broadest sense a register is any logically delimited collection of data, including data sets for purely temporary use. In the context of business statistics we take a more narrow angle:

> *An administrative register is a systematic collection of data which can be related to individual businesses in such a way that updating is possible.*

According to the purpose they serve, administrative registers can be subdivided into basic registers and specialised registers.

- *Basic registers* are maintained as a basic source for public administration in general or for the serving of several different administrations. These registers typically apply to keep stock of the business population and its dynamics. An important condition is that such registers maintain identification attributes also used by other administrations. Moreover they should contain certain basic data of common interest to a number of administrations. Typical examples of basic registers are *administrative* business registers like the French SIRENE and the Danish Central Business Register.
- *Specialised registers* serve one or an explicitly defined limited group of purposes only. These registers are maintained by the authority that is also the user. Basic registers often provide part of the input, such as the basic attributes name, address, legal form, SIC code and size class of legal and local units. Examples of specialised registers are the VAT register and the *statistical* business register.

3. Balancing pro's and contra's

The decision whether to collect directly from businesses or to rely on administrative registers is a matter of balancing the relative advantages and disadvantages.

Main advantages of administrative registers

- avoidance of response burden;
- cheaper than direct collection, in particular when response burden is included as a real cost item in the comparison of costs;
- no or negligible non-response;
- no sampling errors;
- data reported *may* be more accurate because of intensive data checks by administrative authorities.

Main disadvantages of administrative registers

- Discrepancy between administrative concepts and statistical concepts.

  For obvious reasons, concepts regarding variables and units in administrative registers comply to the administrative objectives. In case these differ from statistical needs, and in case the administrative authorities are not able or willing to change or add concepts, either the NSI has to take extra effort to bridge the gap, or accept a devaluation of the statistical concept;
- Poor integration with other data of the statistical system.

  This is in particular a problem when administrative units do not correspond to statistical units, either because their concepts differ, or because of deviating identification numbers. Even if the variables recorded in the administrative register perfectly fit to the needs of the NSI, matching problems can prevent from using them;
- Risks with respect to stability.

  The authorities may change the administrative practices or the legal basis for the register, without taking notice of statistical needs, in particular with respect to disturbing effects on continuity of time series;
- Data reported *may* be less accurate.

  In certain situations, e.g. tax registrations, the respondent has a direct interest with the data he provides. This may cause systematic bias.
- Data may become available with unacceptable delay.
- Legal constraints with respect to access and confidentiality.

Obviously. the pro's and contra's mentioned have no absolute value. It depends of the specific situation whether they apply and to what extent. Therefore, the review has to be seen as a checklist which can be used in the process of decision making.

4. Uses

There are three types of usage of administrative registers:

1. Direct processing.

   This use is justified when the *concepts* recorded fit reasonably well to the statistical system *and* when the *data* supplied are of a sufficient quality. Strictly regulated or supervised activities like banking and insurance or air traffic, which are liable to numerous administrative declarations, will often apply for direct processing.

2. Matching with administrative or statistical sources.

   Such matching may take two forms:

   - horizontal integration, in order to supplement attributes from one register with attributes from another register or to check the validity of the data;
   - vertical integration, to track changes in the stock of units or in their attributes by comparing the registers at two points in time.

   Experiences have learned that in particular the SIC code of a unit is difficult to obtain and to keep up to date in administrative registers. It is, therefore, important that the NSI undertakes a regular and careful check of this item and, where possible, match the administrative register with the statistical business register.

3. Indirect processing.

   This type of use requires processing which is supported by the use of information from a supplementary ad hoc statistical data collection. It applies in case of insufficient coverage of the target population by the administrative register, or in case of units which do not fit to statistical unit concepts.

Example

1. Due to the threshold value as established by the administrative register authorities, the coverage ratio for the variable turnover in the hairdressing industry may be relatively poor. An ad hoc survey or a sample among small businesses can supply additional information, enabling to estimate the missing part.
2. Part of the legal units, maintained by administrative registers, are not active any more, while another part only renders ancillary services to parent units. A special survey, as a follow up on the updates reported by the administrative register, may be needed to trace the inactive units and delete them from the statistical register, and to gather the information needed to join ancillary units with their parents.

5. Summary

To date, administrative registers as an alternative source for direct data collection cannot be ignored. They can be of great help in drastically reducing response burden and surveying costs. Still, once the NSI has decided to make use of such registers, the statistician should be well aware of their shortcomings, both in the field of non-fitting concepts and biased data. He then has to take appropriate measures in order to compensate for the deficiencies and safeguard quality.

II. THE STATISTICAL BUSINESS REGISTER

Author: Ad Willeboordse

1. Introduction

One of the major decisions taken in the output design stage concerned the delineation of the (intended) target population. It was defined as the set of elements to which the statistical data *should* refer in view of the survey objectives.

Making such a statement is one thing, tracing these units in practice is quite another question. In this chapter we will explore the data source most applying for materialising this population, i.e. the Statistical Business Register (SBR). Particular attention will be given to the consequences of the fact that the *frame* population, corresponding with the records in the sampling frame, will in practice inevitably diverge from the target population.

It is important for surveying statisticians to be aware of the three layers which are relevant here:

Example

The *ideal target* population (reflecting user wishes) of the computer services industry in 1996 consists of all KAUs, existing at any moment in 1996 with major activity NACE code 72.

The *target* population (reflecting survey goals) consists of all Enterprises, existing on December 31, 1996 with major activity NACE code 72.

The *frame* population (reflecting reality, as embodied in the business register) of the computer service industry in 1996 consists of all Enterprises, recorded in the Business Register on 31 December 1996 and classified as NACE code 72.

The CR - SBR emphasises the significance of a statistical business register for an NSI by stating that:

> "...business registers for statistical purposes represent a basic element of systems of information on enterprises, making it possible to organise and co-ordinate statistical surveys by providing a sampling base, possibilities of extrapolation and means of monitoring the replies from enterprises, in particular those covered by Directives"

2. The Statistical Business Register and the sampling frame

The ideal situation is the existence of a perfect sampling frame, i.e. an up-to-date set of all elements in the target population. In that case the frame population would correspond fully to the target population. The formation and updating of a sampling frame is, however, a complex and difficult matter. Indeed, characteristics as well as composition of the population are constantly changing over time, while information sources for these changes use to be far from perfect, both with respect to accuracy and up-to-dateness.

For business surveys the most appropriate source to derive the sampling frame from is the Statistical Business Register (SBR), as described in the CR -SBR. This CR obliges the member states to record enterprises, legal units and local units. In a next stage of development KAUs and LKAUs may be added.

Having decided to make use of the SBR, the surveying statistician should be well aware of any differences and shortcomings when compared to the requirements following from the objectives of the survey, as worded in the definition of the target population. In the first place, for most surveys only part of all elements recorded in the SBR will be relevant (if we wish to survey all Enterprises with NACE 72, only a limited number of units apply). Secondly, one has to take into account that the SBR will inevitably contain a number of errors, omissions and delays, so that measures must be taken where appropriate and possible. Therefore, it is recommendable that the surveying statistician has a proper understanding of the contents, the functioning and the shortcomings of the SBR.

3. Contents of the SBR

The SBR as a system

An SBR, maintained by a statistical agency, can be conceived as a *system* transforming data from administrative sources into data, fitted for statistical use. Thus, within the SBR as a system an (administrative) input phase, a conversion process and a (statistical) output phase can be distinguished.

The difference between input and output side of the SBR is also illustrated by the figure shown in PART A-II.5, where different types of statistical units are related with their legal/administrative counterparts.

A Business Register can in practice evoluate to a real *data bank*, containing integrated micro data resulting from a variety of surveys. While recognising that the borderline is not sharp, we will in this chapter confine ourselves to "pure" business registers, i.e. registers containing units and a limited number of attributes, characterising these units.

The output of the SBR

Taking into account that business statistics should together observe and describe a country's total productive activity, the output of the *ideal* SBR can be defined as "an up-to-date file of all statistical units, active within the country's territory and generating value added", as well as their relevant statistical and administrative attributes.

Basic statistical units and basic attributes are considered to be:

- Enterprises and their (institutional) sector, SIC and size attributes;
- KAUs and their SIC and size attributes;
- LKAUs and their SIC, size and region attributes;
- LCUs and their SIC, size and region attributes.

SIC codes may refer to both primary and (one or more) secondary activities.

Other characteristics concern *events* of an economic demographic nature, like birth, death, merger, split off, removal, etc. Furthermore, it is common to record *hierarchical links* between units, like parent-, sister- and daughter relationships, as well as *longitudinal links* between units. The latter enable to track units over time, also when they change their structure.

As updates in the SBR often lag behind real events, it is useful to double *time-stamp* changes by recording both the date of registration and the date of event. The combination of the two gives an idea of the time lag and enables the statistician to eliminate updates which distort time series. Another important aspect is that corrections for errors should be recorded as such, and not as *events* occurred at the date of entering.

Finally, for mailing purposes, information on *names, addresses and telephone numbers* is needed.

In the case of large and complex businesses, which use to fall apart into several statistical units, it makes sense to record the name of a contact person, as well as a detailed description of the part of the business the statistical unit refers to. If the respondent is not able or willing to supply data at the level of the statistical unit, one or more *reporting* units will have to be created and recorded. Separate *collection* units apply when the questionnaires are completed by others, e.g. a book keeping office[i]. If reporting units and contact persons only hold for a particular survey, it might be better to record them outside the SBR in a satellite data base, to be maintained by the surveying unit.

The input of the SBR

At its input side the SBR records units and attributes as supplied by the administrative sources. In most countries these are legal units and their local parts, attributed with names and addresses. For obvious reasons legal units and their local parts have to be linked to their statistical counterparts at the output side.

It is important for an NSI to maintain intensive contact with the administrative data sources, so as to ensure that nature and quality of the information supplied maximally fits in with the needs of the SBR.

Input versus output

The distinction between input and output side of the SBR is primarily of a logical nature. It is meant to stipulate that administrative and statistical units and attributes belong to different worlds, and that there is a gap between them, which has to be bridged by the statistical agency. In practice, in the majority of cases there is a one to one relationship between the statistical units Enterprise res. KAU and the legal unit. The same goes for the relationship between local unit res. LKAU and the affiliate. Still, the exceptions should not be ignored, as these usually refer to large businesses, accounting for a considerable part of total economic performance.

When the SBR is used as the basis for demographic statistics, the distinction between input- and output world is important as well. This is specially true for statistics on birth and death.

Example

Suppose that for fiscal reasons a legal unit, corresponding to an Enterprise, is split up into four legal units: one for each of the functions production, maintenance, sales and book keeping. When statistics are based on the legal/administrative world, this “event” would be accounted for as the death of one (larger) business, and the birth of four (smaller) ones. This could easily lead to the (erroneous) conclusion that young and small units are doing well, whereas large units languish. However, at the level of the economic demography, there is no relevant change at all: one Enterprise remains one and the same Enterprise.

4. Updating sources

Keeping track of the numerous changes in units and their attributes is one of the most challenging tasks of the SBR staff. In this respect it is useful to distinguish between external and internal sources.

- External sources.

  The most commonly used administrative data sources are tax registrations (in particular VAT), registers kept by the Social Insurance Board and by the Chambers of Commerce.

- Internal sources.
  - First, those surveys, using the SBR as a sampling frame, can serve as checks on the existence of units and on the correctness of attributes like SIC and size class. Production statistics seem to be the most fitted for this purpose. Other surveys might be usable as well, in particular when it is possible to include questions relevant for the SBR in their questionnaires.
  - Furthermore, if updates from the administrative sources are found to be of poor quality, follow up actions by the SBR staff may be necessary.
  - Finally, there might be a need for incidental censuses in selected areas.

Large enterprises with complex legal structures and a variety of activities deserve special attention. Improper, outdated or vague unit delineations within these conglomerates may cause considerable errors in statistical reports. Therefore, it is worthwhile to *profile* these dynamic conglomerates frequently, preferably by personal contact. If such operations are not carried out by register staff, it is recommended that the surveying statistician verifies the unit structure of large and complex enterprises as supplied by the SBR, before mailing the questionnaires.

5. Common frame imperfections

There is not one SBR in the world which can claim to provide a complete, perfect and up-to-date representation of any of the target populations applying for business surveys. This implies that the surveying statistician always has to reckon with frame imperfections. Information with respect to the nature and size of imperfections may impose the surveying statistician to take additional actions and measures to compensate for them. As far as this is not possible, their effects can be taken into account when interpreting and presenting the outcomes.

Four broad categories of frame imperfections can be distinguished: undercoverage, overcoverage, multiple listings and misconstruction of units. Practice reveals that there are a lot of reasons for one or more of these anomalies to occur. By means of an example we will list the problems most commonly encountered.

Example

Let us assume the design of a survey, intending to measure turnover of computer service industry in 1995. According to the postulates put forward in Chapter A-III, the target population then can be defined as "all KAUs with principal activity NACE 72 at any moment in 1994". Let us further assume that the sampling frame is selected from the SBR as it shows on January 1, 1995. Now the following discrepancies between the target population and the reality of the specified frame population may occur:

1. *in case the SBR lacks - as is often the case - self employed businesses, this whole category will be missing (undercoverage);*
2. *if units which ceased activities in the course of 1994, would be deleted from the SBR immediately after dying, these will be wrongly omitted from the population (undercoverage);*
3. *for a number of units, the SBR will have missed change of principal activity, taken place before 1994. This means both overcoverage (non-computer service industry units classified as computer service industry) and undercoverage in the frame population;*
4. *a notorious superfluous category are dead units; it is usual for SBRs to contain more than 10% dead units, because external sources fail to trace terminations (overcoverage);*
5. *apart from time lags, there will inevitably be a number of mis-classifications, either caused by external source registers, or by own SBR staff (identified overcoverage and unidentified undercoverage);*
6. *the transformation of legal units to appropriate statistical units like Enterprise and KAU is time and money consuming. Therefore, often legal units are taken as a proxy for both Enterprise and KAU. This will lead to errors for part of the population.*
7. *It is common register practice to often change identification numbers of units, as a consequence of administrative events only. This is problematic for panel surveys, losing units without reason, and for demographic statistics, recording too high birth and death rates.*

6. Coping with frame imperfections

In order to fight the deficiencies mentioned, and thus to reduce the discrepancy between frame population and target population, the imperfections listed can be fought with the following measures:

ad 1.

to fill the gap of self employed businesses, the surveying statistician might consider to make use of one or more additional sources, such as yellow pages or member lists of the trade organisation. This requires the integration of two or more files. In such cases, matching is necessary to prevent multiple listings. This is usually a very labour intensive operation, with only relative success; unit concepts and their attributes are likely to be different. So, while such operations undoubtedly fight undercoverage, they will lead to some overcoverage as well;

ad 2.

this is an example of a situation where register practices may fail to comply with the needs of SBR users. In general it is recommended that SBRs should maintain history, which means that dead units are classified as “non existent” instead of being removed from the frame. Moreover, activity changes should not lead to over-writing of the old codes;

ad 3.

if this anomaly is suspected to occur to a considerable extent, one might consider to expand the boundaries of the frame population, e.g. with units classified in computer service industry-related categories. This will of course lead to a substantial overcoverage in the frame population;

ad 3/4/5

if the total share of delayed updates, misclassifications, dead units and wrongly structured units is considered unjustified high, and if one wishes to avoid the overcoverage just mentioned, selection of the frame population might be preceded by a census to be held in the relevant area and in related areas. This applies in particular for situations where a new survey is set up in an area not covered by other surveys yet. However, the value of such a one-off snapshot of part of the SBR should not be overestimated. It is true that the area checked will be cleared of erroneous units, but the counterpart effect is denied (which is in particular a problem when other SBR areas are not sufficiently covered with regular surveys). Besides, the area will soon be outdated when no regular and comprehensive updating source is available;

ad 6.

as mentioned before, specially large and complex units should be contacted in advance when ties between legal units are not maintained in the SBR;

ad 7.

in order to prevent sample surveys from losing units for administrative reasons only, one might undertake special actions directed towards tracing and subsequently linking of deletions with corresponding registrations.

Whether or not measures as mentioned apply, depends on the relative importance of the anomalies. To be able to judge this, it is recommended that the SBR contains quality indicators for the different areas. At the level of individual units and attributes one should at least record date and source of the most recent update.

Whatever number of such measures may have been taken, a certain amount of errors, missings and delays will inevitably remain in the sampling frame selected. These deficiencies will, at least partly, reveal themselves as *frame errors* during the collection stage.

7. Summary

A statistical business register, being the main source for the selection of sampling frames, can be considered a system in which statistical units are generated out of administrative units. When related to the demands of the target population, a number of imperfections occur. For the survey designer it is important to be aware of these imperfections, and to take compensation measures whenever possible.

III. DEFINING THE SAMPLING FRAME

Author: Ad Willeboordse

1. Introduction

Assuming that the SBR as described in the previous chapter is the main data source for the sampling frame, the definition of the sampling frame involves two important issues to decide on. First the sampling unit must be chosen and secondly the scope of the sampling frame must be delineated. In general this frame is a selection of the total area covered by the SBR.

2. Choice of the sampling unit

In principle the type of units of the sampling frame should not necessarily be equal to the target units of the target population, i.e. to analytical units. There may be even situations where it is more efficient to deviate, and to apply *reporting* units in the frame. However, in most business surveys, such a deviation would be problematical. If, e.g., the target population comprises KAUs and the sampling frame consists of legal units, the respondent would first have to create and subsequently classify his own KAU, before completing the questionnaire. Therefore, in this manual we will assume that sampling frames should preferably be composed of *analytical* units (the situation where, for lack of better, legal units are taken as a proxy for statistical units, does not deny this supposition).

3. Scope of the sampling frame

In an ideal situation, in which the SBR would contain on Jan. 1, 1997 all Enterprises effectively carrying out major activity NACE 72 in 1996, the delineation of the area to cover in the sampling frame would be trivial and straightforward. However, the deficiencies mentioned in par. 3.4 ask for special selection rules. We mentioned already that it might be sensible to include in the sampling frame categories outside of the computer service industry, so as to reduce under-coverage. Another reason for special selection rules relates to continuity of time series.

Example

As far as updates in the SBR during 1996 had the nature of correction for errors or referred to events which took place *before* 1996, one should in principle avoid these *non-real* change (non real in the sense that the change did not take place in the reference year) to affect data in time series. There are, essentially, two ways to handle:

1. the non-real changes are taken into account when specifying the frame population of Jan. 1, 1997, while the population of Jan. 1, 1996 is revised retrospectively. This implies a revision of the data surveyed for 1995 (and possibly previous years), so that the difference between the 1996 and 1995 data reflects real developments;
2. the non-real changes are *ignored* when specifying the frame population on Jan. 1, 1997 (and saved until a revision year).

Obviously, the first method is the better one; the data are more up-to-date, and previous data are revised. However, it takes more effort, because data for previous years have to be recalculated or even recollected. Moreover, neither users nor statisticians like to revise data.

It should be noted that, in a situation where the pattern of lagging is systematic over time, correction operations as described may not be necessary. Of course, the existence of such a pattern must be proven by analysis, before deciding to refrain from corrections. Finally, one should take care to preserve consistency between annual and monthly data

Example

Supposing that monthly surveys of 1996 are based on the frame population per January 1, 1996, the frame population for the annual surveys on 1996 should actually ignore *any* changes in SIC-codes, recorded in the SBR during 1996, be they real or non-real. This method ensures that monthly data aggregate to annual data.

4. Summary

Sampling frames for business statistics will in general comprise statistical units. The borderlines of the frame population cannot simply be derived from those of the target population. When the business register shows imperfections, the first will cover a broader area than the latter.

IV. CHOOSING SAMPLING DESIGN AND ESTIMATION METHOD

Authors: Elly Koeijers and Karin Hilbink

1. Introduction

After defining the target population and the frame population, the next step is the choice of the sampling design and the estimation procedure. Sampling theory provides a variety of methods according to which samples can be drawn and estimates can be produced. The first question is whether to take a sample or to carry out a census. If a sample is considered sufficient then it has to be decided, what size applies. Before sampling it is important to focus on the data to be published.

2. Estimates and parameters

The tables in the final publication contain figures, which we call *estimates*. The term is common in situations where the survey is based on a sample, but if we consider a census to be a (100%) sample as well, all table entries based on observations are estimates. The underlying (unknown) quantities in the target population for which the estimates are given are called *parameters*. Some examples of parameters are population totals, population means, proportions, ratio's of two totals, and differences of two totals. In most surveys, estimates are not only made for the entire population, but also for a number of sub-populations into which the population is divided.

In the design stage of a survey we search for methods to produce estimates lying as near as possible to the values of the parameters. In business surveys it often occurs that large businesses have relatively high values for attributes of interest. If in the survey design no provisions are made for this fact, it is possible that the outcomes of the survey are very inaccurate and no guarantee can be given that the estimates will lie near the values of the (unknown) parameters.

3. Sampling design and estimator

A common starting point in survey design is to concentrate first on the mere effects of sampling and ignore any frame imperfections or any other errors which might occur, such as non-response errors. Although this implies a crude simplification of reality, a clear picture of the 'ideal' survey is very helpful to indicate in later stages where deviations from the starting point might occur. Sometimes research is possible under certain reasonably realistic assumptions, for example concerning the response behaviour of the population. This may yield a picture of the 'best achievable' survey.

When developing a sample survey the following issues are important, assumed that the survey is based on a probability sampling approach:

1. The sampling design and the sample selection scheme, both applying *before* data collection.
2. The estimator by which a particular parameter will be estimated, applying *after* data collection.

The *sampling design* is a set of specifications which define the target population, the sampling units, and the probabilities attached to the possible samples.

Example

The target population consists of all ENT's, existing on December 31, 1996 with major activity NACE code 72.

The sampling unit is the Enterprise.

In case of a simple random sample without replacement of size n, the sample design is defined by the probability of selecting s:

$P(s) = 1/\binom{N}{n}$ for each set s of n different elements from the population and 0 otherwise.

The *sample selection scheme* describes the mechanical selection of a sample according to the chosen design. If the frame population does not entirely cover the target population, in other words if there are frame imperfections as listed earlier, the sample selection scheme operates on the frame population, which implies that one must be careful as regards the definition of the population parameters to be estimated.

Which design fits best for a particular survey, depends on the auxiliary information present in the frame. The more information is available before sampling, the better the sampling design can be tailored to the survey objectives.

Example

When the size of a business is recorded in the business register, this information can be used to measure variables that are related to size. Such variables are labour costs and turnover. The auxiliary information can be used by choosing the inclusion probabilities proportional to the size. A high probability of being selected is attached to large firms.

The *estimator* is the mathematical function by means of which the estimate for a particular parameter is computed. The choice of the estimator is often induced by the form of the parameter.

Example

A ratio of two population totals might be estimated by the ratio of two estimated totals. An estimator may contain auxiliary information, either from the sampling frame or from external sources.

4. Strategies

The combination of a design and an estimator is called a *strategy*. [ii]

In the design stage it is useful to study the performance of different strategies in order to find the best strategy for the survey in question. Before alternative strategies can be compared, one has to know how to handle estimates and their sample-to-sample variability.

For business surveys it is usually relevant to apply the theory of sampling from finite populations, as described in the above mentioned references. Under the conditions for which the theory applies, the stochastic character of the survey results is determined by the chosen strategy. The conditions imply that sampling is carried out according to the sampling design and that the observed data are true (i.e. no measurement errors), while no non-response occurs and no errors are introduced during the processing of the data.

The stochastic character of the survey results can be understood in the following way. The sample selection scheme can be implemented by means of a computer program. The idea is that if the program would be run repeatedly, then 'in the long run' each possible sample would come up with a frequency corresponding to the probability defined in the sampling design. Sample-to-sample variability refers to the fact that different samples give rise to different estimates for a particular parameter, assumed that for each sample the characteristics of the selected elements are observed under the same general conditions.

To study the performance of a strategy, we have to consider, at least conceptually, all possible estimates that might be generated by the combination of the sampling design and the estimator. Important aspects of a strategy are the bias, the variance, and the mean square error of the estimator.

5. Bias

The concept of bias requires a way of thinking in terms of all possible outcomes of a sample survey. Bias is not connected with one particular estimate, but relates to all estimates for a certain parameter the sample survey might produce. The idea of bias is easier explained, if we first consider the concept of expectation.

The *expectation* of the estimator can be considered as the average of all estimates generated 'in the long run' over repeated sampling and subsequent observation of the characteristics of interest. When the design poorly fits the estimator, the combination of the sampling design and the estimator will possibly generate poor estimates, far away from the values of the parameters in question. In the distribution of the estimator the expectation acts as the centre of gravity, which does not always coincide with the value of the (unknown) parameter.

The *bias* is the difference between the expectation and the value of the population parameter. If the bias equals zero, then the estimator is *unbiased*. Bias might occur, e.g., when the survey population and the frame population do not entirely overlap. Another cause of bias might be a bad combination of the sampling design and the estimator. Non response might yield bias as well.

Example

Suppose we want to measure turnover of businesses in the population. We choose the probabilities proportional to the size of the business, as proposed in the previous example.

A poor combination of a sampling design and an estimator might occur, because there is a lack in the communication between the person who is responsible for the sampling design and the person who works on the estimation. The surveyor who is responsible for the estimation phase is not aware of the fact that the sample was drawn proportional to size and thinks that every business has the same chance of being selected in the sample. He attaches the same weight to each element. This will lead to an overestimation of the total turnover, because the contribution of the large businesses in the estimator is larger than it should be regarding the sampling design. In other words: this estimator is biased.

Unbiasedness is a desirable property for an estimator, because it guarantees that over repeated sampling the results are not systematically too low or too high, although the highest and the lowest possible outcome may lie far apart from each other.

6. Variance and mean square error

Measures which describe variability are the *variance* and the *mean square error* of the estimator. The variance can be considered as the 'long run' average of all squared differences between the estimates and the expectation of the estimator. Likewise, the mean square error averages the squared differences between the estimates and the parameter. It can be easily shown that the mean square error is equal to the sum of the variance and the squared bias.

The variance expresses how close the estimates lie near the expectation of the estimator, whereas the mean square error measures the closeness around the parameter. The term *accuracy* usually refers to the variance and *precision* to the mean square error. An estimator is called *accurate* if the variance is small and *precise* if the mean square error is small. In most surveys one cannot draw all possible samples to calculate the variance, but in many circumstances the value can be estimated on the basis of one single sample.

For unbiased (or nearly unbiased) estimators the accuracy is often expressed in terms of the *coefficient of variation*, which is defined as the square root of the estimated variance divided by the estimate of the parameter. An advantage is that the coefficient of variation is a dimensionless quantity.

Example

If the variable of interest, for example turnover, is measured in dollars, the variance has a dimension dollars^2. In many surveys it is difficult to relate to the often large number of the variance. In the computation of the coefficient of variation the dimension vanishes and a percentage that is easy to interpret, results.

7. Comparing strategies

The performance of a strategy can be considered on the basis of various aspects, i.e. biases, variances, mean square errors of the estimators involved. A problem is how to compare alternative strategies in the design stage. How can the performance of various strategies be judged before the actual survey is carried out?

The situation is relatively simple when relevant information is available for all elements of the entire population. This is the case when a survey is planned to be continued on a sample basis, whereas it was previously carried out as a *census*. In this case variances, etc. can be computed, either by mathematical expressions or by simulation experiments. A problem may be that the data are somewhat outdated, but the information is complete for the entire population.

Example

To shorten the time needed to conduct a survey in which every element used to be examined, one can decide to switch from a census to a sample. The information of the last census can be utilised when different strategies are considered. Variances and standard deviations can be computed and used when the effects of sampling designs and estimators are examined. In the interpretation of the results one has to bear in mind that the data used can be somewhat outdated.

The situation is more complex when the available data are based on a *previous sample*. This sample has been drawn according to one sampling design, whereas the performance of different alternative strategies should be compared. Variances, etc. have to estimated, which can be problematic. Again, the information can be outdated.

Example

Suppose we want to redesign an annual survey that used to be conducted on a sample basis. In this case the only information available is the data of the respondents of the survey of previous years. Variances and standard deviations of the variables of interest can be estimated from the data of last year and used when examining the effects of sampling designs and estimators. Great care has to be taken when using these estimated variances. Because these variances are based on a sample they can deviate considerable from the true variance of the population of last year. The problem that the data can be somewhat outdated also applies in this situation.

When *no data* are available at all, a pilot survey might be conducted to gather information on the characteristics of the population. After the pilot survey the above mentioned situation occurs, in the understanding that the data will generally be more up-to-date. It must be emphasised that the size of the pilot survey should be chosen not too small, otherwise the results may be useless.

8. Sample size

Once the combination of sampling design and estimator being decided on, the sample size can be determined. Two aspects may play a role: cost and precision. Usually the precision will increase if the sample size increases. However, the larger the sample the more expensive and time consuming the survey will be. Often the sample size is decided by the *total budget*. In some cases *precision* is a more important factor than cost. For the determination of the adequate sample size a linear cost function can be a useful tool. The procedure for stratified random sampling is described in Särndal, Swensson, and Wretman (1992).

9. Stratified sampling

Introduction

In business surveys the method of *stratified sampling* is widely used. Prior to sampling, the population is divided into non-overlapping sub-populations, called *strata*. The strata can be treated as separate populations for which suitable strategies can be chosen. The selection of the samples in each of the strata is carried out independently.

Example

To enhance the precision of the estimation of the total turnover of computer service firms it is possible to stratify the population into size classes. We can stratify the population in three strata of size class, which is measured as the number of employees in the business:

The size class *small* consists of businesses where at most 9 persons are employed, the businesses in size class *medium* employ 10 to 99 persons and the businesses in size class *large* 100 or more.

The sample is allocated over these strata and for every stratum a sampling design is defined before the survey is conducted. The samples in the three strata are independent.

Reasons for stratification

The first step in stratified sampling concerns the choice of the characteristics by which the strata are to be formed. This choice depends on the purpose for which stratified sampling is applied. Some reasons for stratification are:

- enlarge precision (at a given sample size);
- produce estimates with a specified precision for separate strata or for sub-populations consisting of more than one stratum;
- control fieldwork efficiently;
- use different frames for different parts of the population.

Precision

If more precision is the reason for stratification, it is beneficial to form strata which are more or less homogeneous groups in the sense of the target variables. In business surveys size class usually does well as stratifying variable, because size is often highly correlated with most variables of interest. It is then necessary that the size attribute is available for all businesses in the sampling frame. Almost inevitably the information on size is somewhat outdated, but it remains useful if size is not too unstable through time.

Allocation

Another issue is the way in which the total sample size is allocated to each of the strata. In business surveys the allocation is generally chosen disproportionate over size classes. If a statistician has the choice between the inclusion of either a large or a small business in the sample, then it commonly turns out that the gain in overall precision is more substantial if the large business is preferred to the smaller one, as the latter accounts fro a smaller share in the total performance.

On the other hand, if the collection cost for the large business is higher than for the small one, then the total cost of the survey increases less if the small business is chosen instead of the large one.

Optimum allocation

The problem of *optimum allocation* can be considered from two different viewpoints, precision and cost. If the overall precision is fixed, then the sample size can be allocated to the strata in a way that the total cost attains a minimum. Conversely, if the total cost is fixed the overall precision can be maximised.

Special cases of optimum allocation procedures are the method of *the Neyman-allocation*, the *x-optimal allocation* and the *power allocation*. For details we refer to Cochran (1977) and Särndal, Swensson, and Wretman (1992). The best allocation for one target variable is likely to be not optimal for another. Possible compromises are *average variance*-methods and the *convex programming*-methods, for which we refer to Bethel (1989).

The determination of an optimum allocation is often an iterative process. The first step may yield in some strata a sample size larger than the corresponding number of businesses in the population. The usual procedure is to take all businesses in those strata and subsequently reduce the total sample size or the total cost with the contribution thereto of the businesses already included. Then the procedure is repeated for the remaining strata.

10. Domain estimation

Sub-populations for which separate point estimates are required are called *domains*. Domains can be characterised, for example, by region. A domain can be any sub-population or the entire population. Unless domains can be defined as separate strata, the sample size cannot be controlled for each domain and varies randomly from sample to sample. If the number of elements in the sample within a domain is a random variable, then the precision deserves special attention. More on domain estimation is given in chapter IX.

11. Surveys repeated across time

In a statistical agency most business surveys are repeated, usually on a monthly, quarterly or annual basis. This adds a time dimension to the choice of the sampling design and the estimation method. In sampling across time two extreme options are:

- drawing a new sample on each occasion;
- retaining the same sample throughout all occasions.

Between these two extremes a variety of designs applies. Kalton and Citro (1993) distinguish the following types:

1. Repeated cross-sectional survey (a new sample on each occasion).
2. Panel survey (the same sample throughout all occasions).
3. Repeated panel survey (a series of panel surveys, overlapping or non-overlapping, aimed at longitudinal analysis of gross changes).
4. Rotating panel survey (repeated panel survey with overlap, aimed at both estimating cross-sectional totals and net changes).
5. Overlapping survey (series of cross-sectional surveys with maximum overlap taking into account both the changes desired in the selection probabilities for sample elements that remain in the population as well as changes in the composition of the population over time).
6. Split panel survey (combination of a panel survey and a repeated survey).

Type 5 is particularly useful in business surveys, as stated in Kalton and Citro (1993), because the desired selection probabilities may vary from one occasion to the other, so as to reflect changes in size and activity. Some designs are inappropriate for certain objectives. Panel surveys, for example, ignore changes in composition and characteristics of the population due to inflow, because the sample is drawn from the population existing at the start of the panel.

It depends on the objectives of the survey, which type of design is most suitable. Roughly spoken, the following guidelines can be given:

- Both for net and gross change, it is advisable to retain as much as possible the same sample throughout all occasions.
- For a current cross-sectional total it is relevant that the sample is drawn from the current population.
- For averaging a series of cross-sectional totals it is advisable to draw for each cross-sectional total a new sample from the corresponding current population.

When a survey across time is planned, it is necessary to define the target population over time. The definition must contain at least one or more time points or a specified period of time. Some relevant questions to be decided on deliberately before starting the survey, are:

- Should the population consist of all businesses active at a given time point?
- Or is a period of time more important and is any business active during that period a member of the population?
- Should the definition include those who were only active temporarily?

References

Bethel J. (1989), Sample Allocation in Multivariate Surveys, Survey Methodology, 15,1.

Cochran, W.G. (1977), Sampling Techniques, Wiley, New York.

Hedayat, A.S. and Sinha, B.K. (1991), Design and Inference in Finite Population Sampling, Wiley, New York.

Kalton, G. and Citro, C., (1993), Panel Surveys: Adding the Fourth Dimension, Survey Methodology 19, 2, p. 205-215.

Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M.P. (1989), Panel Surveys, Wiley, New York.

Särndal, C.E., Swensson, B., and Wretman, J. (1992), Model Assisted Survey Sampling, Springer-Verlag.

V. QUESTIONNAIRE DESIGN

Authors: Hans Akkerboom and Jean Ritzen

1. Introduction

The content of the questionnaire can be considered as a balance between two possibly conflicting interests, i.e. information supply and demand. Information demand is expressed by the information needs of the users, as eventually specified in the tabulation outline. Information supply is determined by the data suppliers' willingness and capability to provide data. Supply and demand have to be matched in various phases of the survey design process. Again, we emphasise the iterative nature of the survey design process. Sometimes, what happens in the 'questionnaire design phase' may have consequences for decisions already taken in the output design phase.

In attaining the balance more and more priority is given to the capability to provide data by respondents. Specially in the case of quantitative data which must be derived from the business accounts it has to be understood that the items in the questionnaire must fit the data available. Similarly, in the processing and analysis phase various forms of quality control may reveal shortcomings in the survey design process which have to be taken into account in forthcoming redesign efforts. Ideally, the survey process is subject to continuous reengineering, in which a balance has to be struck between flexibility and continuity. In particular, the development of a questionnaire is an iterative process. How and how often one iterates between exploration, design and evaluation depends on innumerable considerations of quality and costs.

2. Consultation of respondents

The first step

Consultation of potential respondents in the initial stage of questionnaire design is the first step to match information supply to information demand after translation and standardising the user needs inventoried in a well defined statistical output.

To summarise the previous chapters, in this step general survey content and information exchange procedures (feasibility studies) are fitted to respondent's needs and capabilities, so as to optimise 'general survey conditions'. These are reflected in data communication conditions, such as:

- way of conferring the information request to the sampled units,
- reminder strategy,
- data transfer medium (electronic, mail, telephone, fax, voice recognition, touch-tone data entry),
- data transfer mode (self-administered or interviewer-assisted);
- informed consent,
- guarantees of confidentiality and/or anonymity (disclosure control),
- available interviewers or contact persons.

The second step

The second step to match information demand to information supply is in questionnaire design. Here the main aim is to reduce the risk of various measurement errors by constructing a user-friendly measurement instrument, given the general survey conditions. User-friendly means that users can easily understand what data are asked for, in which format, to what accuracy and to what purpose, and that they can provide these data with no more cost and effort than necessary.

Consultation of (volunteer) respondents should at least consist of

- a 'qualitative pre-test' to detect and amend major bugs in the questionnaire ('detailed tryout' with a few respondents),
- a 'quantitative pilot' to detect and amend more subtle problems in the questionnaire ('general tryout' with many respondents).

The final questionnaire prototype should be tested as part of the whole survey process in

- 'a grand rehearsal' to get the whole conduct of the survey going.

The third step

The third step to match information demand to information supply is in quality control during conduct of the survey, mainly by:

- data quality inspections on subsamples (re-interview methods, e.g. in the form of 'response analysis surveys', in which respondents are followed up to obtain quantitative information about accuracy, consistency and other quality aspects of the data supplied);
- debriefing of interviewers, debriefing of staff involved in editing and checking questionnaires or debriefing of respondents (either quantitative by debriefing questionnaires, or qualitatively by debriefing interviews or experiences, individually or in groups);
- preparation of adequate documentation of procedures, etc. for operations staff undertaking the collection;
- preparation and implementation of training sessions for operations staff.

Measurement errors discovered during this step will in general only be corrected if they can be accounted for in the editing process during the processing phase. While editing is traditionally considered as a data correction tool, it can also be considered a form of data quality control aimed merely at error detection. Thus, response analysis surveys and editing information serve both the documentation of data quality to the users and the suggestion of redesign options.

3. Data communication demands

While consulting respondents in the questionnaire design phase, information about conditions on the communication process will naturally emerge, especially as regards willingness and capability to provide data.

Willingness to provide data depends to a large extent on the way the information request is conferred, on the time and effort needed to provide the data, and on the value and use of the data as perceived by the respondent. These items should be checked in an early stage, i.e. during the *qualitative pre-test*. This includes respondent consultation about use and interpretation of introductory letter, explanatory material (brochures, notes) and follow-up procedures (help-desk, reminders, recalls).

The introductory letter is a decisive factor in obtaining the co-operation of the respondent. It should be brief and convincing, and it should comprise at least information on:

- the goal of the survey;
- its mandatory or voluntary status;
- how confidentiality is ensured;
- which department or person may be the most suitable for completing the form
- expected average completion time;
- name and telephone number of contact person in the NSI.

Also, the comfort of use of data transfer materials should be checked (envelope to return a form, floppies, electronic medium, etc.).

In general, the information request should be convincing and conveniently timed. It should be backed by relevant trade organisations, preferably by letting them send a letter of recommendation to respondents. If possible, summaries of data collected earlier should accompany the request. Both information provided earlier by the same respondent, and tables of earlier output relevant to the trade and/or region in question apply.

The possibility of providing the required data will largely depend on the accounting practices and management information systems in use with the respondent, but also on the clearness and conciseness of the information request. An important practical point to be checked is that the person to which the information request is directed is indeed the right one to organise the data collection within the sampled unit.

4. From output variables to questionnaire concepts

Questionnaire design amounts to matching information supply and demand at a micro level. This requires transformation or break-up of output concepts into ‘manageable data items’, that is data separately available in management information or accounting systems. To reduce response burden, derivation of the output variables from these items should be done as much as possible by the NSI.

In this chapter, the data items will be called *questions*, irrespective of their occurrence in electronic, paper or other format. The total set of items makes up the *questionnaire*. EDI is a new and promising tool to marry the worlds of supplying respondents and demanding statisticians. We will devote a separate chapter to this topical issue, which is considered one of the most powerful weapons in fighting the form filling burden.

While operationalizing output concepts into *questions* one has to be aware of measurement errors the respondent may make when completing the form. It is important to distinguish three types of errors:

- we speak of *omission* when the respondent fails to answer a question because he does not notice it, or because he deliberately avoids it or because he does not understand it;
- conversely, we speak of *commission* when a respondent gives information he was not asked for. This can arise as a consequence of misunderstanding or incorrect assumptions;
- we speak of a *mistake* when a respondent gives incorrect information for any other reason.

Commissions are easier to notice than omissions. Mistakes are not always readily identified during the collection phase. In that case they should be picked up during observation studies or during the form testing stage.

5. Design of a prototype questionnaire

In business surveys, a questionnaire, defined as a set of manageable data items, is traditionally associated with a form, but, as mentioned above, how a questionnaire looks like depends on the method of data collection (medium and mode). It is important to have the questionnaire available for various media (e.g. both for mail, either by paper or by floppy, and for telephone) and for various modes (e.g. both for self-administration and for interviewer-assisted data collection). This makes it possible to let the respondent decide on medium and mode most suitable to his needs and possibilities.

While designing a prototype questionnaire to be tested in a pre-test and/or pilot, experience with form design and data quality issues is a valuable asset. The proof of the pudding is in the eating, so that a pre-test and pilot with real-life respondents is necessary each time a new survey is mounted (or redesigned). It is well-known that even with unchanged formulations of data-items, interpretations and connotations change over time and depend on the changing organisation of the enterprise and its accounts. Ideally, any question should be tested anew in any new context.

Indeed, the main concerns in form design (formulations and layout, either on paper or on a computer screen) are

- ample introduction: the why, how and when of data supply;
- ease of interpretation: clear, understandable and concise;
- motivation: interesting, attractive and readable;
- guidance (flow): the respondent is guided clearly and logically from section to section and from question to question.

Similar considerations apply to the introductory letter and any other accompanying materials. No ‘rules of questionnaire design’ are generally valid, but in general it may be helpful to obey the following ones:

1. Avoid abbreviations.
2. Be as specific as possible and feasible; specify dimensions, required accuracy and detail, etc.
3. Be specific about the reference frame: time and place, units, things to be excluded or included. If possible provide the respondent with pre-print unit-identifying information on the questionnaire. If not ask for information which is suitable for checking the interpretation of the unit, for example the number of persons employed.
4. Avoid double-barrelled questions (two in one).
5. Closed questions are preferred above open questions if possible.
6. Provide separate sections (and the whole form) with clear titles.
7. Make clear where to find instructions.
8. Put explanatory notes and instructions as close as possible to where they may be needed (as close to the question as possible).
9. Use layout to enhance salient concepts, instructions, etc.
10. Use layout for a clear ‘picture’ of what is required.

11. Avoid as much as possible questions that are not relevant as perceived by the respondent and avoid other redundancy.
12. Rank the questions in a logical way.

How the rules are put into in practice depends on the method used for the data collection. The rules can vary, depending of the data collection mode applied.

6. Collection media

Once the content and wording of the "questionnaire" has been finalised a decision is required on the most appropriate collection medium to obtain the required information from the respondent. The principal methods available are:

- mail enumeration;
- telephone (including the use of Computer Assisted Telephone Interview
- (CATI) technology);
- personal interview (including the use of Computer Assisted Personal
- Interview (CAPI) technology;
- electronic data capture (by means of tape, diskette, or electronic data interchange (EDI).

The selection of the most appropriate collection medium involves consideration of a number of issues. These include the relative cost of each methodology against available resources (including skill levels), the complexity of the data items being collected, and the amount of time respondents normally require to provide/extract the information from their records (which is largely a function of the number and type of variables being collected).

The primary advantage of *mail enumeration* is that it is normally the least costly (per respondent) of the main collection methodologies. It is therefore suitable where a collection includes a large number of respondents, and where respondents require considerable time to extract the required information from their financial records. The primary disadvantages of this methodology includes the time and resources required (through reminder action) to elicit acceptable response rates, and the amount of query action required to obtain the correct information after the questionnaire has been processed/edited. Both factors may lead to delays in the publication of results. Both factors may also be alleviated through the use of sound form design standards and adequate forms testing.

Obtaining the information *by telephone* has the advantages of reduced lead times in the collection of information, and of being able to query the respondent at the same time, if the initial response is incorrect. This methodology is most suited for collecting very small amounts of information which the respondent can provide at the time of the telephone call. In this respect it is suitable for collecting key sub-annual (e.g. monthly, quarterly) indicator information (e.g. total retail turnover, total number of employees, etc.). The application of computer technology with this approach facilitates consistent real-time query action (enabling information to be compared with data previously obtained from the same respondent or with industry averages, etc.), and the data entry of correct/edited information.

The collection of information by *personal interview* enables complex data items/variables to be obtained from respondents. The quality of the information collected is largely a function of the skill of the interviewer, as well as forms testing, etc. The interviewer is able to probe and query the respondent at the time the information is obtained. The reliability of this method makes it particularly useful for sample surveys where the collection of accurate information from a relatively small number of respondents is essential. It is also useful where a speedy response is required. The major disadvantage of this approach is the high per respondent cost. The use of CAPI technology also facilitates the application of consistent real-time query action.

The primary advantage of capturing information from respondents *electronically* (EDI) is the possible reduction of respondent load. This medium is most suitable where the content of the "electronic questionnaire" is relatively stable over time (with respect to the variables being collected). Such stability is essential to enable respondents (if necessary) to program their systems to extract the required data. The use of electronic data capture may require the development of the electronic questionnaire prior to the commencement of the reference period. EDI is considered one of the most promising tools for both reduction of response burden, decrease collection and processing cost and increase product quality. Therefore we will discuss this issue in more detail in a separate chapter.

7. Questionnaire Design Standards

The systematic improvement of questionnaire design practices in an organisation is facilitated by the documentation of a clear set of mandatory standards and guidelines outlining best practice (such as those outlined above (in par 4 of Section 5 on page 59)). The main elements of such documentation include:

- any mandatory standards with respect to cover design, legal obligations, font, layout, etc.;
- standard definitions of key variables which staff designing questionnaires have to use. The purpose is to reduce unnecessary variations in the definition/specification of variables. Such standardisation facilitates the comparison of data across different statistical collections, and the development of databases;
- clear statements of mandatory forms testing procedures and principles that staff developing questionnaires have to use before a new or modified form can be used in a "live" collection.

8. Testing procedures

As a result of the preceding phases, the statistical office has concluded that the content of the (first prototype) questionnaire (including the explanatory notes and the introductory letter) meets the survey objectives . Make sure that any remaining issues about survey content and its operationalization are clearly stated. For the moment, then, subject matter experts give word to the (volunteer) respondents in informing about whether they can complete the questionnaire in the desired way at a reasonable 'cost' (time and effort). Questionnaire design experts should act as intermediaries in the testing process: they discuss pre-test plans and pilot plans with subject matter experts. Questionnaire design experts should be sufficiently (but not partially) informed about the general survey conditions: aim and general content, design parameters, data communication conditions. Various procedures may be invoked to test the questionnaire: in any case, one has to involve a broad range of respondents reflecting different types of measurement problems that are suspected to occur (self-employed enterprises as compared to multi-activity enterprises, service firms as compared to production firms, etc.) Testing possibilities are (repeated) forms of

1. Document design analysis.
2. Questionnaire design experts score the questionnaire content and layout according to certain checklists (cf. rules 1 to 10 above). Among other things, the check items may relate to problems previously encountered in qualitative pre-tests or quantitative pilots.
3. Informal 'test interviews' with respondents.
4. The questionnaire is completed by various respondents (say 20) and the questionnaire design experts observe and code what happens. After completion, questionnaire design experts may ask why certain things happened and how. Observation of respondents may reveal many practical problems if done 'on site', that is at the relevant business. Observation in a 'cognitive laboratory' may be a practical way if individual test interviews have to be followed by a group interview in a 'focus group discussion'.
5. Qualitative pretesting by 'focus group discussions' with respondents.
6. A group of comparable respondents (size of firm, main activity, etc.) is invited to discuss the prototype questionnaire in detail in a meeting at a convenient site (at least somewhere where observation through one-way screens and/or audio registration is possible). The participants in the focus group are asked to be very critical. They have to be informed about the prototype questionnaire in some way or another. It can be sent by mail beforehand, or it can be presented and completed just before the group discussion starts, as appropriate. Usually consulted respondents do not restrict comments to questionnaire items, but will review the general data communication conditions salient to them (say with respect to data availability, questionnaire flow, response burden). Group size may be somewhere between 5 and 10.
7. Other evaluation methods.
8. Debriefing questionnaires are to be completed by interviewers or respondents that had experimental experience with the prototype questionnaire. Debriefing of 'independent subject matter experts' not involved in the survey design process. Debriefing of technical experts on form layout, printing, desktop interviewing, etc.
9. Qualitative pretesting by cognitive methods with respondents.
10. A few (say 20 to 30) respondents are invited to take part in in-depth investigations of the process of completing the questionnaire. Preferably this is done 'on site' and otherwise it may be done in some 'cognitive laboratory'. Various techniques are used to let the respondent be a 'meta-respondent' as well: he is stimulated to inform about his own considerations and decisions while completing the questionnaire, apart from providing the data themselves. 'Thinking aloud' and 'paraphrasing of key concepts or whole questions are two familiar techniques. Cognitive methods may reveal subtle problems with interpretation, information processing and selection of an eligible answer. Thus they may suggest changes in wording, answer format, inclusion and exclusion criteria, specification of reference period, etc.

Apart from debriefing questionnaires, consultation of (volunteer) respondents typically involves between 50 and 100 enterprises, specially by combining a few informal test interviews and focus groups with a series of cognitive interviews. These are attractive ways of consulting because they are effective and cheap for detection of measurement problems. Cognitive methods require more work but can be very rewarding. Beyond detection of measurement problems, cognitive methods easily and cheaply provide explanations of and possible solutions to measurement problems.

9. Pilot survey

In a quantitative pilot study, not only the questionnaire is to be tested but also the method of data collection and the data processing. A pilot involves a sample of the target group. The number can be between 100 and 800. Such a size allows quantitative conclusions to be drawn about a few alternatives for the questionnaire. Methods for 'design and analysis of experiments' and debriefing of interviewers are useful instruments to improve questionnaire and process.

The pilot survey is meant to give a clear insight in cost and efforts of data collection and data quality. This is the basis for final decisions on the general survey conditions and on the basic question whether to proceed with conducting the survey or not. A ‘grand rehearsal’ of the survey may be carried out to solve any unexpected, remaining operational problems. The decision to have such a rehearsal will in general follow the decision on whether to conduct the survey or not.

Experiences from the pilot survey will not only lead to final modifications in the questionnaire or other general survey conditions, but they may also lead to a reduction of the level of output ambition, both in terms of contents and scope. As a consequence, the intended tables outline may be reduced as well.

10. Summary

Like the table outline is meant to serve the needs of the users, so the questionnaire should satisfy the “needs” of respondents. It does not require much imagination to characterise these needs: speak the language of the respondent, minimise the number of questions and comply maximal with his accounting and management information system. Developments like EDI promise excellent chances to meet these conditions. In any case, taking these conditions serious inevitably implies a certain information gap between survey yields and table outline needs. This gap has to be bridged by the surveying statistician during the processing stage.

PART C SAMPLING, DATA COLLECTION AND DATA ENTRY

In the previous stage we "produced" the tools with which the data collection is carried out: we designed the questionnaire and the sample selection scheme. And we described the ins and outs of the sampling frame as embodied in the Statistical Business Register. In this part we will put these tools into use:

- With the help of the sample selection scheme the sample is actually drawn from the sampling frame.
- Questionnaires are then sent to the sampled units and - partly - returned by respondents.
- The data collection stage ends with entering the data reported in a data file, comprising the records of all sampled units.

It is in the stage of data collection that the relation between surveying statistician and respondent becomes apparent. This relation is presently dominated by three issues, which provide one of the major challenges of the years '90:

- reduce response burden to a minimum;
- reduce non-response rates to a minimum;
- make maximum use of advanced technologies, in particular EDI.

These highly related issues, which determine the core of the corporate surveying strategy, will be dealt with first.

I. MINIMIZING RESPONSE BURDEN

Author: Ad Willeboordse

1. Introduction

National Statistical Offices throughout the world face the need to drastically reduce response burden as one of their major strategic challenges for the years '90. As the design of a new survey unavoidable increases this burden, there is extra reason to stress this topic in this handbook. Besides, minimising response burden is not only in the interest of respondents. A successful reduction program may very well reward the statistician too: data quality can be higher, response will be quicker and response rates will rise, while collection cost fall.

Apart from this all, caring for respondents appears to become a condition to survive. After all, respondents are the wells out of which surveyors quench their thirst for data. If we don't take care of them, the well will dry up.

The concept of "burden" has two dimensions, a quantitative and a qualitative one. The first refers to cost in terms of time and money and the second to perception. A response burden policy should explicitly involve the latter aspect, because in the end it is perception which is, more than workload, decisive for the willingness of respondents to co-operate.

When considering ways to reduce or minimise response burden, a distinction can be made between a corporate strategy, concerning the package of business surveys as a whole, and measures taken at the level of individual surveys. For both a number of attention points will be considered. But first we will clarify the concept of response burden, so as to avoid misunderstanding with respect to the meaning of this key issue in modern survey policy.

2. The concept of Response Burden

As is the case with many statistical terms, the concept of response burden is interpreted in various ways. We present a number of alternative definitions in the form of dichotomies. When discussing the issue, and in particular when considering to measure response burden, it is important to make sure which concept applies. This implies that for each of the pairs a choice must be made.

- *Objective versus subjective burden*

Whereas objective response burden directly refers to the cost respondents have to make for completion of questionnaires, subjective burden reflects their perception. Which of the two burdens is the "heaviest" depends to a large extent on the idea the respondent has of the usefulness of the statistics resulting from his efforts. The distinction is in particular relevant when one compares the response burden of large and smaller businesses. While the latter carry a much larger objective burden, the first use to be the heaviest complainers. Their subjective burden is higher, because they often do not make use of statistical data.

- *Gross versus net burden*

This distinction applies when one wants to quantify response burden. While net (objective) burden takes into account the "benefits" enjoyed by respondents for their contribution, gross response burden denies them. The most tangible feature of benefits is the feedback of survey results to the respondent.

- *Imposed vs. accepted burden*

Imposed response burden assumes that *all* respondents sampled will *fully* and *consciously* complete the questionnaire with sufficiently *accurate* data. Accepted response burden takes a more realistic approach: only the *responding* businesses are accounted for, at the *real* completion

cost. This means *i.a.* that questionnaires filled in by or with the help of field service, do not or only partly generate response burden. It also means that increasing non response rates have a favourable effect on the response burden

- *Maximalistic vs. minimalistic concept of burden.*

Form filling demands more from the respondent than just fill in a questionnaire. It includes a number of diverse acts, varying from opening the envelope and reading the introductory letter to responding to follow-up calls. Thus, there may be a considerable difference with burden figures based on pure completion time

Monitoring Response Burden

Many NSI's develop burden reduction programs, either voluntary or imposed by government. These programs often comprise targets: the response burden has to be reduced with x % in y years time. Such quantitative targets require estimates. There are, however, more justifications to quantify and follow response burden over time. Response burden can be considered as a cost component in the process of making statistics, and it is only self-evident that a (survey) management information system keeps track of *all* costs. Information on respondents cost can help survey managers in balancing cost and benefits of surveying in general, and in considering alternative resources like administrative registers as well.

Data on response burden can be obtained either by estimates or by direct measurement. Estimates (mostly by survey management or field staff of the NSI)) have the advantage of avoiding extra response burden (sic!) but are of a subjective nature. Government and business organisations may easily reject figures estimated by the NSI, claiming that this highly interested party cannot be objective "by definition".

Direct measurement requires adding a question to each survey form, or during follow-up actions. For obvious reasons, such a question should be extremely simple. Therefore, it is better to ask completion *time* than completion *cost*, although the latter is conceptually superior.

Example

The following question could be added to the survey questionnaire:

Please provide an indication of the total time it took your organisation to complete this questionnaire. Do not confine yourself to the filling in as such, but take into account **all** *activities needed to complete, like:*

reading and understanding the accompanying letter and explanatory notes;

gathering and processing of data taken from your records;

gathering of additional information.

Total completion time: ***..... hours ..... minutes***

Special care should be given to the explanatory note. This should be worded such that the respondent perceives the question as being in his own interest rather than an extra burden.

Example

Such a justification could go as follows:

We are fully aware of the fact that filling in this questionnaire imposes a certain burden. Therefore, we take ongoing effort in reducing response burden. In order to monitor the results of these efforts, we would like to know how much time it took you to complete this form.

Taking into account the various concepts as mentioned above, the question arises which choices apply when monitoring response burden, either by estimates or by direct measurement. Although the rule should be that "different concepts (apply) for different purposes", in most circumstances the following choices will be preferred:

- *objective* burden. Subjective burden is in some respects more relevant (e.g. as a measure for acceptance and willingness to cooperate), but is much more difficult to measure;
- *gross* burden. Net burden would require the quantification of the value of data published, which is even more difficult. Moreover, this value would differ per respondent;
- *accepted* burden, because it is more realistic than imposed burden. Still, for internal NSI uses, there is one disadvantage: increasing Non Response rates have a positive effect on response burden. To avoid such undesirable "rewards" and consequently a less alert attitude towards declining response rates, survey managers should be confronted with burden figures which include hypothetical non response burden as well.
- *maximalistic* concept of burden, simply because it is the more realistic one.

Response burden reduction rates should be judged cautiously. Such performances must always be considered in relation with other performance indicators like user satisfaction and cost-effectiveness.

Indeed, changes in response burden can be induced by a variety of causes, sometimes of a very different nature. A reduction of the burden as a consequence of the termination of a statistic is quite another "performance" than a reduction induced by the introduction of EDI, given an equal or even lower level of cost and an equal or even better output performance. Therefore, one should try to break down the changes in response burden into categories which give more insight in the measures behind them and in the effects on output and collection- and processing cost.

3. Response burden policies applying at corporate level

A corporate surveying strategy assumes the presence of a centre of interest within the NSI, where one has an overview of all business surveys and the domains they cover. Besides, for effective knowledge and control one should have disposal of a central data base indicating for each individual business which surveys it is actually involved in.

There are several instruments for carrying out a corporate strategy, directed towards reduction of response burden.

- Co-ordination, concentration or integration of data collection.
- Respondents generally dislike to receive numbers of questionnaires from different departments within an NSI. Moreover, when surveying departments act more or less independently, there is a risk of overlaps and redundancy in questionnaires. Obviously the most annoying situation occurs when a respondent is confronted with different definitions for the same data-item.
- For these reasons alone, individual surveys should be developed within the context of a co-ordinated survey strategy. Such a strategy aims at minimising the number of questionnaires and contact points. For a new survey this means that first one has to consider whether the information need can be met by incorporating the questions in an existing survey. Also when this is not possible, the number of contact points can be further reduced by concentration of surveying activities in such a way that one particular respondent only communicates with one department within the statistical bureau. Of course, there may be good reasons to deviate from this ideal construction. Still, even when respondents are tackled from different places in the organisation, contacts can be streamlined by appointing an *account manager* who is responsible for a harmonised approach of a particular (group of) respondents. Integration of questionnaires and clustering of surveys may not only reduce (perception of) burden, but also contributes to the consistency of reported data and thus to quality of statistics.

Example

The variable “number of employees” is one of the most frequently collected data-items in business surveys. As a consequence, one respondent has to report several times on several forms about one and the same item. Moreover, as definitions will most likely differ among questionnaires, he cannot suffice with mere duplication: he has to calculate the answer repeatedly. On the other hand, the NSI will provide different and inconsistent figures on the number of employees.

All three interested parties will suffer:

1. respondents are irritated and annoyed
2. NSI faces needless processing cost
3. users get confused.

Possible measures are:

1. integration of surveys/questionnaires. In case this measure is considered too rigorous:
2. concentration of surveys (one counter policy). In case this measure is considered too rigorous:
3. elimination of all questions on “number of employees” but one. In case this measure is considered too rigorous:
4. attuning of concepts/definitions. In case this measure is considered too rigorous:
5. confrontation and attuning of data from different surveys before publication. It should be noted that there is one legitimate reason to persist to redundant questions on (particularly) number of employees: check whether the respondent reports about the appropriate unit. The latter is specially important in case of complex unit structures.

- Co-ordinated delimitation of sampling frames.

  Survey populations are often delineated according to field of economic activity: there are separate production statistics for separate SIC groupings. Care should be taken that a particular respondent is not classified in different groupings at a time. This risk can be reduced by drawing samples for all such surveys from one unequivocal source, i.e. a centrally maintained business register. However, a true guarantee for avoiding overlaps demands more: the different surveys covering different SIC areas should apply the same type of statistical unit, as well as a uniform *method* and *moment* of determining their respective sampling frames from the business register.

Example

Assume two surveys: the production statistics for 1995 on electronic industry and on computer service industry. The first is based on enterprises, the latter on KAUs.

1. Assume a business with two activities: manufacturing of computers and Y. In the Business Register the business is recognised as one Enterprise, consisting of two KAUs.

Effect: KAU Y will be an element in of both surveys. The damage is twofold: double form filling burden for the respondent and double counting in statistical results.

2. Assume the sample for X is drawn from the sampling frame as it existed in the Business register on December 31, 1995, while the sample for Y is based on the situation on January 1, 1996.

Effect: Enterprises/KAUs that changed activity from X to Y are double counted; those who changed from Y to X are denied.

Notice that the anomaly is not caused by different moments of drawing the *sample*. It may very well be that a survey draws the sample in June 1996 or even a year later. The problem is caused by deviating *sampling frames*.

- Co-ordinated sampling.

  Control of response burden can be achieved by co-ordinated selection of samples. Without any internal co-ordination within the statistical agency it might happen that some businesses receive more forms than others, although these businesses are comparable in size, activity, etc. A powerful tool to spread the response burden is the combination of a centrally maintained business register and a comprehensive computer program for co-ordinated sampling.

- EDI.

  Electronic tapping of data directly from respondents accounting systems is a very promising way to reduce burden. Making use of EDI enforces surveying statisticians not only to comply with accounting practices, it also stimulates centralisation of data collection operations, thus leading to the advantages mentioned above. The use of EDI techniques is further explained in CHAPTER III of This Part.

- Information on response burden.

  Although quite a challenge, NSIs should aim at informing respondents in advance about the surveys they will be involved in. Ideally, the NSI should send a comprehensive listing of these surveys, including average completion time of the questionnaire, at the beginning of each year. Of course, such a frank attitude is only possible with a very well planned and

centrally organised surveying strategy, while all of the above mentioned issues should be realised or at least underway. Although a long listing might deter respondents, the opposite effect is also possible. Burden perception surveys among respondents have revealed that they tend to ascribe surveys of all kinds of other institutes to the NSI, so that, so that the list might even arouse a positive surprise. Moreover, one might consider to add a listing of surveys for which they will be *excluded* from sampling in the current year, so as to prove that the spreading of response burden really works.

4. Policies applying at the level of individual surveys

The final product of a survey is the result of the combined efforts of respondents and statisticians. The policy should be to design the survey such that the share of the respondent is limited to the necessary minimum. Central issue in this policy is the notion that the intended statistical output does not automatically determine what the respondents input should be. This is true for all the elements of the final product, as specified in PART A. Given the contents and the quality standards of the output, a number of measures apply.

- Number of respondents.

  This is the most trivial aspect and the message is simple: take samples as small as possible. The use of advanced sampling techniques and high quality sampling frames contributes to this goals. Besides, one should make maximum use of auxiliary information, which might be derived from other surveys or administrative sources. Respondent friendly behaviour requires also that the survey sample strategy is flexible enough to leave room for varying sample sizes over time.

Example

- In case of a monthly survey. one might vary sample sizes in such a way that any third month the sample is higher while in between lower sizes or even cut offs suffice. Quality main remain acceptable by, e.g., interpolation.
- In case of an on-going panel survey smaller units might be delete from the sample by rotation.

- *Units.*

  A first demand is that the observation unit be defined such that the respondent can recognise himself as a real transactor in the economy rather than an artificial construction. This can be realised by stressing the requirements of *autonomy* and *data availability* in operational unit definitions, while accepting a certain degree of *heterogeneity* . The KAU, as defined in the CR -SU is an outstanding example of such a “real life” unit. There will, however, always remain situations where the respondent is not able or willing to report data at the level of the statistical unit envisaged. Then the *reporting* unit will deviate from the statistical unit. In case of a 1:*n* relationship the statistician will have to allocate the data reported, while in the inverted case consolidation applies.

- *Variables: concepts and definitions.*

  Whenever statistical concepts deviate from accounting concepts, this should be considered a problem for the statistician and not for the respondent. This means that questionnaires should be designed in such a way that they can be completed directly from book keeping records, and that it is, again, up to the statistician to bridge the gap between questionnaire concepts and statistical output concepts.

- *Variables: number and detail.*

Like sample sizes, it might be justified to alternate the contents of questionnaires. Once the "maximum" questionnaire is designed, one should seriously consider whether it is really necessary to apply it *full size* for *each* respondent during *each* reporting period. This goes in particular for small firms, where big detail is often overdone. Whether it is wise to vary number and detail of variables over time, e.g. by asking detail only at intervals of three months, depends on the situation. In certain cases it is better to maintain a constant rhythm. This is e.g. the case when respondents take special measures to generate the data required automatically. Besides, there is the advantage of the *learning effect*.

- *Variables: accuracy.*

  For smaller units the burden may be relieved by collecting data in ranges rather than discrete values, without notable effect on the quality of statistical data.

- *Tailor-made questionnaires.*

  When a survey covers distinct SIC-areas, accounting practices and vocabulary may differ among branches. This may require different questionnaires for different groups of respondents.

- *Relevance of questions and explanatory notes.*

  Response burden is not only determined by the time it takes to *answer* questions, but also by the time and effort needed to *read and understand* questions, introductory letters and explanatory notes. These should not only be brief and clear but, above all, applicable. Not or hardly applicable questions and extensive irrelevant explanations are extremely annoying for respondents. This is another justification for tailoring questionnaires to homogeneous groups of respondents. SIC-code and size class are applicable attributes to identify such groups, as well as data reported in related or previous surveys. Pre-printing of product specifications is an example of the use of the latter. A probate way to fight overkill is *double sampling*. This requires the conduct of two consecutive surveys. First a few simple questions are asked to a large number of enterprises, e.g. whether or not a business carries out R&D and if so, whether the expenditures exceed a certain threshold. The findings are used to narrow the population for the second stage. Besides, it is possible that the first stage generates data which can be used as estimators in the second stage, enabling the statistician to reduce the sample size.

- *Feedback of results.*

  Providing respondents with the results of the survey is a measure which, of course, only affects *perception* of burden. The information to supply should be carefully composed and one must make sure that it is of real interest to the businessman. Otherwise, the effect might as well be negative.

- *Measure respondent load.*

  In section 2 we explained the usefulness of on-going monitoring of the response load. We also noticed that the addition of a question regarding time spent on form completion *might* have a positive effect on perception of burden. If properly introduced, respondents may consider such a question as an indication that the NSI is aware of their problems and tries to do something about it. Besides, outliers might be given after-care by advising them how to reduce the completion time.

5. Summary

Minimising response burden should be an issue to take into account during all stages of the survey design process. A number of measures, norms and attention points apply, some of which ask for a corporate strategy towards data collection. It is true that in some respects the easing of reporting conditions induces increasing efforts for the statistician, but there are advantages as well, both with respect to efficiency and data quality. A smart sampling design and a respondent friendly questionnaire are powerful tools in the struggle against overburden. However, prior to this all, the possibilities to make use of existing administrative registers must be deliberately considered.

References

De Vries, W.F.M., Keller, W.J., and Willeboordse, A.J., 1996, Fighting the Form-filling Burden, International Statistical Review.

II. MINIMIZING NON-RESPONSE

Author: Ad Willeboordse

1. Introduction

For obvious reasons, the issue of non-response is closely related to that of response burden. One might even reason that the degree of non-response is, to a certain extent, an indicator for the burden as it is perceived by the business world. The growing response burden of the seventies and the eighties has been accompanied by falling response rates in the whole range of business surveys all over the world.

The close link between response burden and response rates implies that many of the measures, directed towards reduction of the burden, will have a positive effect on response rates as well.

In this chapter we will not repeat these measures and policies. We will confine ourselves to listing a number of specific actions deliberately aiming at a direct increase of response. The measures recommended are partly based on experience with a "real life" survey conducted by Statistics Netherlands. First, we have to make sure what we exactly mean by the concept of (non-) response.

2. The concept

Like with response burden, there is a variety of interpretations of the concept of response. The deviations stem from different views on how to deal with:

1. returned questionnaires which are only partly filled in;
2. questionnaires, completed partly or entirely by field service;
3. questionnaires, returned after the deadline for processing.

Whether or not to incorporate these borderline cases in the response rate is primarily a matter if usage of the rate. If it means to be an indication for the willingness of respondents to cooperate, the second category should not be included. For such a purpose it might be interesting to make a further breakdown of the rates in spontaneous response and response induced by follow up actions.

In case the response rate is used as a quality indicator, the second category should be included while the third should be left out. This is the figure that should appear in publications as part of the explanatory notes for the user.

The first category poses a borderline problem for which no concrete rules can be given. Actually, it points to a weakness the whole concept of response suffers from: ideally, the response rate should be based on counts of the questionnaires returned after weighing *all* of them with the quality of the answers. This is, of course, not feasible in practice; still it is important to be aware of this weakness when interpreting the response rates.

As a general rule we can state that partly completed questionnaires should only be counted as response if they are *usable* in the processing stage.

It is recommended to document response behaviour in such a way that all of the categories can be recognised.

3. Measures to increase response rates

Example

Most of the measures recommended below stem from an action, undertaken by Statistics Netherlands to reduce non-response in the quarterly paper survey on "construction works in progress ". The action resulted in an increase of the response rate from 65 to 95 %.

Measures, already mentioned in the context of response burden reduction, like compliance with accounting standards, will not be repeated here.

- dispatch questionnaires around the date accounting data are supposed to be available. A few days earlier is no problem; but avoid delays of more than a week;
- the final date of return should not be too far away (for quarterly surveys not more than two weeks) so as to avoid respondents to forget the questionnaire;
- send first reminders within a week after final return date;
- telephone reminders give higher scores than written reminders, but are more expensive. Therefore, start with one or two written and end with a telephone reminder;
- give priority to larger businesses, specially with respect to expensive actions, like personal visits;
- permit guestimates, not as a "general pardon" but on an individual basis, i.e. for those respondents who are willing but not able to deliver the data in time;
- always contact respondents - and do it quickly after return - who only partly complete questionnaires or report implausible data (e.g. always the same numbers), so as to avoid the impression that their behaviour does not matter;
- invite respondents to contact the Office in case of questions and make sure that there is always a qualified staff member to answer the telephone;
- take care that the mailing list is based on up to date information, so as to avoid mis-addressing;
- try to make appointments with contact persons and mention name and room number on the address label;
- reward respondents with (extracts of) publications. If they are not interested - which is likely the case with accounting staff - let them indicate other staff within the organisation;
- stimulate survey staff by providing them with response score (and comparisons) on a continuing basis;
- train staff in handling difficult respondents;

In case of statutory surveys, the question arises when to institute legal proceedings. Prosecution is to e considered the last resort. General guidelines cannot be given, but in most cases a reserved policy applies. One might, e.g., confine prosecution to those respondents *consistently* refusing to co-operate with *any* survey. Full ignorance of the legal possibilities would make the NSI inveracious, whereas overkill might lead to negative publicity.

III. EDI AS A TOOL FOR DATA COLLECTION

Authors: Ad Willeboordse and Winfried Ypma

1. Introduction

Electronic Data Interchange is one of the exposures of advanced information technology which will inevitably have a dramatic impact on the statistical process. In this Handbook we will discuss EDI primarily as a tool for data *collection.* However, we will stress and illustrate that the changes in the data collection phase of the process will affect the subsequent stages of the statistical process as well.

Given the present state of affairs of most business surveys, i.e. data collection by paper questionnaires, we will discuss the EDI issue from the angle of survey *re*design.

2. The basics of EDI

The basic idea behind EDI in the data collection process is quite simple: given the fact that both respondents and the NSI hold electronic information systems, why not establish a direct electronic link between the two systems? As usual, the answer is less simple than this rhetoric question suggests. *Direct* communication between the two systems may be hampered by:

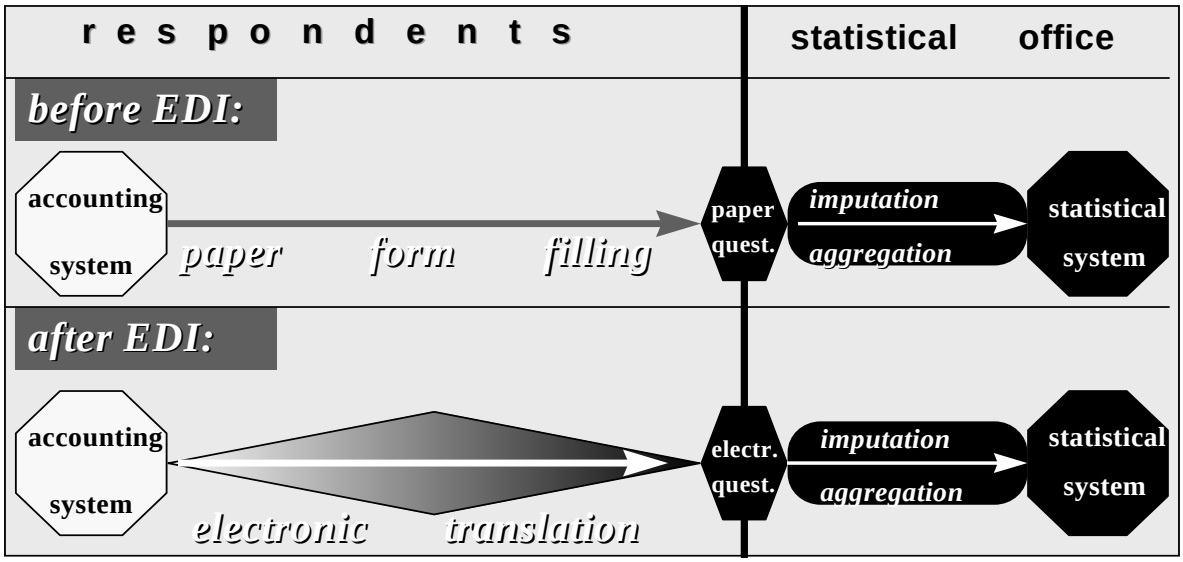
- technological dissimilarities, such as data structures, formats and software, as well as limited data communication;
- conceptual dissimilarities, such as:
  - ⇒ naming and coding of data items;
  - ⇒ level of aggregation of data items (a statistical item me be composed of different accounting items) ;
  - ⇒ existence of data items (a statistical concept may have no accounting counterparts).

This situation holds for the major part of the business surveys belonging to the European statistical system. Therefore, edification of data collection is not confined to automation of paper questionnaires, but comprises the design of an electronic *translation* facility as well, so as to bridge the technological and conceptual gap between the worlds of respondents and NSI.

Following the nature of the dissimilarities as mentioned above, the act of translation may refer to four different aspects:

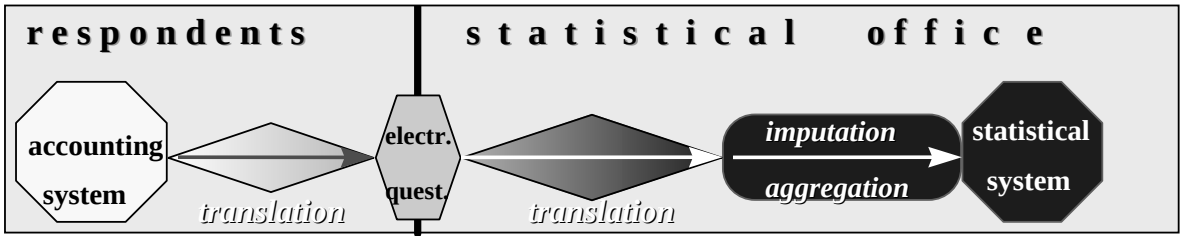
- a *technological* aspect, i.e. transformation of data structure and format from one system to another;
- a *linguistic* aspect, i.e. the translation of names and codes from the language of the respondent to the language of the questionnaire;
- a *computational* aspect, which applies when the questionnaire data item has to be *calculated* from a number of accounting (sub-) items, e.g. by addition and/or subtraction;
- an *estimation* aspect, referring to questionnaire items for which there are no pure equivalents or building blocks in the respondents records (e.g. 1 : n or m : n relations). The data have to be estimated on the basis of nearby or related items, on which conversion keys are applied.

The latter three aspects actually represent the explanatory notes from the paper questionnaire, in the understanding that after edification these notes are expressed in terms of the business accounts language and in automatic translation rules. The move from paper to electronic can be pictured as follows:



This figure pictures the situation where the statistical contents of the electronic questionnaire fully correspond to those of the paper questionnaire it replaces, under the assumption that it is indeed possible to bridge the gap electronically. In practice, however, this assumption often only holds after adaptation of questionnaire concepts in the direction of accounting concepts. When the adaptation would amount to 100%, (electronic) translation by the respondent no longer applies. Provided that the statistical system has to remain unchanged, this implies that the translation from accounting concepts to statistical concepts must be done by the statistical office itself (we leave aside whether or not this translation occurs electronically):

A typical edification project, like we will discuss in this paper, often shows a mix of both extremes:



Notice that the statistical system has shifted slightly to the left; it is indeed not unusual that some concessions are done towards accounting concepts. The figure illustrates in what respects the scope of edification projects tends to reach beyond information technology: the contents of the questionnaire are affected, as well as the statistical processing and the statistical system.

In summary, edification projects, while primarily oriented towards reduction of response burden, tend to imply a process of critical reconsideration with respect to the questions whether:

- users really *need* what they use to get. Obviously, this question holds in particular for data which do not show in business accounts. Putting forward this question is not mere form. The growing tendency to compile "reality-statistics" in stead of artificial constructs, has created a favourable climate for adaptation towards the perception of businesses as actors in the economic processes;
- we should really *collect* all the data we need. Calculation, imputation or estimation, either on the basis of nearby or related business accounting concepts, or obtained by confrontation with data from other surveys or by statistical integration, can be a worthwhile alternative.

Next to this critical reflection on the *contents* of the questionnaire, a third question arises with respect to the *organisation* of data collections. The boundaries between surveys - and consequently questionnaires - are traditionally delineated without taking into account the way accounting systems are organised within businesses. EDI more or less forces surveying statisticians to tune their surveying strategy to the latter. On the one hand, EDI logically demands *integration* of those different questionnaires, drawing on one and the same (sub-) accounting system. This leads to banishing one of the most irritating manifestations of response burden, i.e. redundancy, caused by overlapping questionnaires. On the other hand, EDI requires *segregation* of questionnaires for the completion of which data have to be taken from different (sub-) accounting systems.

3. Basic EDI modes

The nature and complexity of edification projects will vary among surveys. The most relevant discriminating factors are in this respect:

- the distance between business accounting systems and information needs with respect to the technological and conceptual dissimilarities as mentioned above;
- the degree of standardisation of business accounting practices.

When applying these criteria, three main modes of EDI emerge:

A. *direct tapping.* There is no translation needed at the collection side. Edification contains the selection of the relevant items of the business accounts and installation of communication provisions. This mode applies when *demand* of the statistical office - as reflected by the questionnaire - fully corresponds with business accounts *supply*, which logically implies that all businesses use the same accounting system, both technically and conceptually. These ideal conditions seldom occur in practice;

B. *standardised translation.* One electronic questionnaire is designed, as well as one or a limited number of standard translation modules. This mode applies when there is a moderate distance between demand and supply and when accounting practices are highly standardised;

C. *unique translation.* In case of non-standardised business accounts, each respondent has to establish its own unique translation module.

Although strictly spoken not satisfying the definition of EDI, it is nevertheless useful to add a fourth category:

D. *semi-EDI.* The (technological or conceptual) distance between accounting and questionnaire concepts remains too large to be bridged by automatic translation: the electronic questionnaire is filled in manually.

With respect to the modes A and B, a further distinction applies by introducing the *type of respondent* as an additional criteria:

1. administrative bodies, keeping registers for other than statistical purposes.
2. book keeping offices, representing clusters of businesses;
3. individual businesses.

The sequence of this listing is not coincidentally: the higher the ranking, the more favourable the impact on response burden. Therefore, edification projects should in principle aim at attaining the highest possible level in the list. The first mentioned possibility has been discussed in an earlier chapter, so we will focus on the others.

4. EDI and response burden

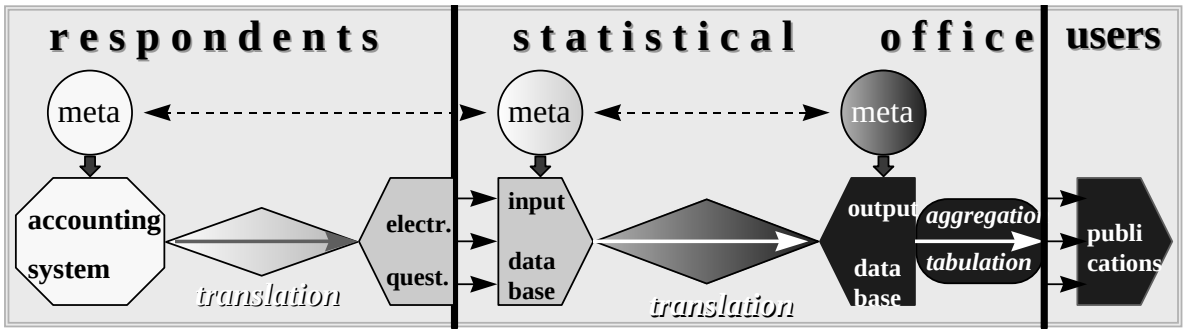
The EDI mode ranked as A will bring response burden practically back to zero. For mode B, it stands to reason that the NSI supplies standard software packages to the respondents. If these are provided for free here too the burden will practically disappear. When mode C applies, the lack of standardisation makes it unavoidable that the respondent himself does the job. This burden is, of course, of quite a different nature than the paper burden it replaces. Recurrent efforts for questionnaire completion are replaced by initial investment to implement the translation rules, followed by - probably minor - recurrent maintenance cost. The complexity of the rules and the stability of the accounting practices ánd of the statistical system will determine the size of the remaining response burden. Although difficult to estimate in advance, there is good reason to expect that the remaining burden will amount to a fraction of the paper burden only.

5. EDI and meta data

The figures show that after edification the data flow from accounting to statistical systems will become significantly more disciplined. In order to establish reliable translation rules, the respondent must be informed precisely about the meaning of the data items he has to report on. Therefore, definitions should leave no room for deviating interpretation; questionnaire items must be defined exhaustively, i.e. in terms of *inclusions* and *exclusions* of items from the book keeping records. Moreover, there may be a need for a variety of questionnaire types, each of which is designed to fit in with the language of a specific, homogeneous group of respondents.

These complex relationships on the one hand, and the growing need for *coherent* statistical data over the whole range of business statistics on the other hand, ask for a highly disciplined and co-ordinated processing of data. A central *input meta data base*, listing all relevant concepts and their definitions, is an indispensable tool to effectively manage and control these processes.

Actually, *two* meta data bases apply: next to the *input* meta data base there is a need for an *output* meta data base as well. While the former defines *questionnaire* concepts in terms of accounting language, the latter in turn defines *publication* concepts in terms of questionnaire items. Where the first supports the respondent in establishing “his” translation rules, the latter supports the statistician in setting the rules for translation of input data to output data.



Thus, translation does not only apply for the respondent, but for the NSI as well. Each of the two parties takes its share in bridging the gap. The choice of the questionnaire concepts eventually determines how the total burden is distributed over respondent and statistical office. The act of designing the questionnaire can therefore be seen as a trade off. Obviously, it should be NSI policy to account for the largest part and to leave to the respondents only what is easily convertible in automatic rules. This policy goes in particular for the *estimation* aspect of the translation. Here, response burden is not the only reason why statisticians should prefer to do the job themselves: the possibility of *control* is a strong argument as well. However, this policy should not be taken in an absolute sense: if a specific translation activity can be done *better* and *cheaper* by the respondent, this is a good reason to leave the work to him.

References

Ad Willeboordse and Winfried Ypma, EDI-The Electronic Matchmaker, Paper presented at the 5th meeting of the IAOS, Reykjavik, June 1996

IV. DRAWING THE SAMPLE

Author: Ad Willeboordse

1. Introduction

Drawing the sample comes down to application of the sampling selection scheme on the sampling frame as defined in the previous part. This drawing for a particular survey is pre-eminently a *co-ordinated* action. Indeed, the operations to be undertaken for each separate survey should be part of a corporate strategy applying for the whole of business surveys. As we pointed out earlier, such a strategy is in the interest of both respondents and users. The first benefit from less and equally spread response burden, and the second from more consistent data.

Assuming that the sample is drawn from the SBR, the first action to take is selection of the relevant part of the register, i.e. the actual delineation of the sampling frame. Subsequently, the sample is taken from this frame. Equal distribution of response burden is the main reason for this operation to be carried out in a co-ordinated way as well.

2. Delineating the sampling frame

While discussing the contents of the business register, we stressed the fact that surveying statisticians will have to live with and be aware of a number of imperfections, when they delineate the sampling frame. In other words: in practice, it seldom occurs that the sampling frame is a true representation of the *survey population* looked for.

Example

The SBR of country X comprises 1.500 computer service enterprises on January 1, 1996. Because of the common frame imperfections, as listed in the chapter on statistical business registers, the real number and composition of the computer services industry at that date will most likely be different. There is, however, no information about the degree of deviation.

The choice of the boundaries of the sampling frame depends to a considerable extent of the severity of shortcomings of the SBR. This goes in particular for the delineation of attributes like SIC code and size class.

Example (continued)

Assume the SBR is used for the production statistics on computer services industry in 1995, which comprises all enterprises with 20 and more employees.

Knowing that:

- the SBR updates are lagging to a certain extent;
- the computer service industry is a booming industry;
- there is a tendency for firms engaged in other business services to change their activities to computer service industry;

it may be wise not to confine the delineation of the sampling frame to enterprises classified with NACE 72, but to expand the frame with units in lower size classes and "neighbour" SIC codes.

If the survey is a cut-off and covers a certain SIC category, and if both characteristics are lagging behind, neighbouring classes may have to be included in the sampling frame.

Ideally, the whole business population must be included in the sampling frame, because in theory any unit might have changed its activity to computer services. In practice, one has to find an acceptable balance between cost and benefits. Moreover, at least part of the other activity areas of the business population will be covered by other surveys. When these surveys generate information on activities (like is the case with product specifications), they may serve as an additional source of information as well.

3. Coordinated delineation of the sampling frame

Taking into account that the number of business surveys may amount to more than a hundred, it is important that the total business population as recorded in the SBR is allocated over these surveys in a co-ordinated way. By this we mean that:

- there may not be a chance for a particular enterprise to receive questionnaires for two surveys on the same subject (e.g. for capital investment in both manufacturing and transport industry). The reasons for this requirement are twofold: avoid needless and irritating response burden and avoid double counting in statistics;
- there may not be a chance for a particular enterprise to belong to different sub-populations for different surveys covering different subjects. When, for example, a business is classified in transport industry for the employment survey 1996, while it is classified in manufacturing industry for the investment survey 1996, users cannot compare employment and investment data.

Ensuring co-ordination requires a highly disciplined and centralised policy with respect to deline-ation of sampling frames. Its establishment could therefore best be left to a central co-ordination unit rather than the individual survey managers themselves.

Example

Assume the (annual) survey on R&D in manufacturing industry for 1996 is mailed in January 1997, while the capital formation survey in manufacturing industry dispatches in April 1997. Co-ordination requires that both surveys draw their samples from one and the same sampling frame, i.e. that of 31 December 1996.

This example illustrates that co-ordination has its price: the April survey cannot benefit from updates, referring to events in 1995 and, because of the time lag in tracing changes, recorded in the SBR between January and April. Of course, as always in statistical practice, the co-ordination rule should not be taken too absolute. If a considerable part of these updates have the character of corrections for errors or delays, one might prefer quality to co-ordination.

Another dimension of co-ordination relates to consistency between short term and annual data. In this respect special measures apply in case attributes like SIC code and size class are only updated once a year, i.e. on January 1. This is common practice in order to prevent discontinuities in monthly time series and in order to prevent businesses from receiving two questionnaires for annual statistics, each referring to a part of the year.

Example

Assume that for a particular enterprise the SBR records an SIC change from manufacturing to transport industry on September 1, 1997. It is best practice to "save" the change until January 1, 1998, so as to avoid that the business has to complete two forms for annual surveys on 1997, one for the period before September 1 and one for the rest of the year. Saving the update until January 1, 1998 rather than December 31, 1997 implies that the annual statistics for 1997 consider the unit as a *manufacturing* enterprise. This is necessary in order to ensure co-ordination with sub-annual surveys for 1997, which have already submitted data for this enterprise in the manufacturing industry.

4. Co-ordinated sampling

Once the sampling frame being identified, the sample can be drawn according to the sampling design and the sampling selection scheme, following from this design. Already in this stage, weights can be attributed to the sampled units.

A form of control of response burden can be achieved by co-ordinated selection of samples. Without any internal co-ordination within the statistical agency it might happen that some businesses receive more questionnaires than others, although these businesses are comparable in size, activity, etc. A powerful tool to spread the response burden is the combination of a centrally maintained business register and a comprehensive computer program for sampling.

Several statistical agencies in Europe apply co-ordination of sampling for business surveys. The method used at INSEE in France is described by Cotton and Hesse (1992). For a description of the EDS-system of Statistics Netherlands we refer to Van Huis, Koeijers, and De Ree (1994). At Statistics Sweden the SAMU-system is in use, see Ohlsson (1992). These three systems operate

on random numbers assigned to the businesses in the register, although the methods differ. The French system manipulates the random numbers after each selection, whereas the random numbers in the Dutch and Swedish systems are permanent.

It should be noted that, for methodological reasons, notorious non-respondents cannot be "punished" for their non-cooperative behaviour: they are equally treated as respondents. Sampling is purely a matter of applying predetermined probabilities. It would not be good practice to allow these probabilities to be affected by their response behaviour. Therefore, control of response burden in the selection stage cannot go beyond the co-ordination of sampling.

Although spreading of response burden does not reduce *total* non response, it certainly has a positive effect on *perception* of burden by respondents, specially when they are explicitly informed about the system. In this respect the ultimate manifestation of a conscious response burden strategy would be to send respondents at the beginning of each year an overview of surveys for which they will be sampled, as well as of surveys for which they will be spared.

References

Cotton, F. and Hesse, C. (1992), Co-ordinated Selection of Stratified Samples. Proceedings of Statistics Canada Symposium 92 Design and Analysis of Longitudinal Surveys, 1992.

Van Huis, L.T., Koeijers, C.A.J., and De Ree, J. (1994), Response Burden and Co-ordinated Sampling for Economic Surveys. Netherlands Official Statistics, 4, 1994.

Ohlsson. E. (1992), SAMU, The System for Co-ordination of Samples from the Business Register at Statistics Sweden. R&D Report Statistics Sweden, 1992:18.

V. DATA COLLECTION

Author: Ad Willeboordse

1. Introduction

Once the units to sample being known, preparations for questionnaire mailing can start, as well as the creation of a file containing records with questionnaire items for each of the units to survey. During the collection period, follow up operations occur in order to enhance response and data quality. Most of the operations can be supported by modern automation tools. After questionnaire dispatch, the main concern of the surveyor is to stimulate and assist respondents, in such a way that response rates and quality of responses are high enough meet the standards.

2. Who is who

When discussing statistical units, we introduced several categories of units, among which *reporting* units and *observation* units. While the last category applies in the context of statistical output, the first one is only relevant in the stage of collection. A reporting unit refers to an address, which is important for labelling the questionnaire. The data filled in on the questionaire relate to the observation unit. If reporting unit and observation unit do not coincide, and the business has a complex juridical or physical structure, it is of utmost importance to clearly specify on the questionnaire about which part of the organisation the respondent is supposed to report. It is here that *legal* units become important. Although not having the status of statistical units, they can be of help in making clear to the respondents where to report on and where not.

3. Automation tools

In this stage of the statistical process, automation tools can be of substantial help. Of course, the ultimate in this respect is the complete “edification” of questionnaires, resulting in a direct link between statistical (input) data bases and accounting systems. But advanced technology can be useful in the case of paper forms as well. Starting from the (paper) questionnaire and a file of respondents, automation tools apply for the following operations:

Identification

Addresses and identification data can be printed on the questionnaire directly, or on a separate self-adhesive label, to be attached to the questionnaire. Thus there will be less chance of missing or unreadable identification data on the returned forms than when leaving it up to the respondent to fill in these data.

Bar coding

Identification codes can be printed as bar codes and scanned back from the collected forms. Best results are obtained by using special printers and appropriate software. Hands free scanners, switched between keyboard and computer, are to be preferred because they can easily be applied in any program.

Use of software tools

Addresses derived from a business register have a highly structured nature, so a database program is best suited for survey and follow up management. Most database programs have excellent form printing facilities. Some of them can also act as a front end to a larger database system, e.g. Oracle. When addresses have to be included in letters, use can be made of a mail merge option, featured by many text editors and word processors.

Automation also creates possibilities for tuning the contents of questionnaires maximal to the specific situation of the respondent. This applies e.g. for product specifications and deleting certain questions or detailed break downs for smaller units.

The information can be derived from micro data sets, both containing data from previous periods and from other surveys.

4. Setting up a micro data file

The list of sampled units and the questionnaire items provide the ingredients for the setting up of a micro data file, in which the data to survey will be recorded. The elements of the set consist of *reporting units*, while the fields of the records refer to *questionnaire concepts*.

5. Checks before dispatch

We noticed earlier that the sampling frame will inevitably suffer from a number of shortcomings, some types of which will be reflected in the sample as well. So, after preparing the collection control files and before making the questionnaire labels, a number of clerical and computer checks apply. Among the most important are the removal of obvious duplications and updating of recently reported address changes, as well as checks on the unit structure as registered in the SBR. The latter goes specially for large and complex businesses, whose legal and operational structure often changes.

In general it is certainly worthwhile to spend some time and efforts at checks and last minute updates prior to dispatch. They prevent irritations with respondents, they help reduce respondent burden and non-response and they save time in later phases of the collection process.

6. Pre-announcement

Respondents should be informed in advance in case of new surveys, as well as substantially changed questionnaires. This gives them the opportunity to adapt their information systems. If possible the pre-announcement should be accompanied by a recommendation of the trade organisation or other institutions representing respondents. Besides, the NSI can promote the survey by release of information to the media, outlining the survey objectives.

7. Help desk

Respondents should be emphatically invited to contact the statistical bureau in case of any problems. Problems, commonly put forward by telephone, refer to unit identification (as

described in the next section), to requests for delay or dispensation and to requests for field service assistance . In planning help desk capacity, one should take into account that immediately after mailing the questionnaires, a large number of such calls will occur.

8. Unit identification

In a dynamic business world, unit structures and activities will frequently change. All types of events, like merger, split off and deconcentration may occur. Conglomerates are often so dynamic in this respect that it is hardly possible to follow them in the SBR, even in the presence of a separate Large Business Profiling unit. Discrepancies between the reporting unit envisaged and actual reality, as it shows up during data collection, may therefore be expected. Surveyors should be alert and signal possible changes to profilers or SBR staff. Failure to detect changes in structure might give rise to serious distortions in the outcomes.

When encountering such a discrepancy, it is important to establish whether it is caused by a recent change, occurred after the reference period, or by an error or lag in the business register. In the first case, special arrangements with the respondent apply in order to reconstruct the former the latter case, actually a *frame error* has been detected, and the reporting unit should be adjusted correspondingly. Whether this also applies at the level of the corresponding *statistical* unit, depends on the policy with respect to the treatment of non-real changes and with respect to co-ordination between short term and annual statistics. Moreover, in case of a sample survey complications may arise in the stage of weighting .

Corrections and updates of units and their attributes should take place in close co-operation with SBR staff. Arrangements have to be made so as to avoid double work and, even worse, double contacting of respondents.

VI. DATA ENTRY

Author: Frank van de Pol

1. Introduction

Different methods apply with respect to entering the data collected in the micro data file as created in the previous stage. Also, decisions have to be taken with respect to the timing of data entering and the inclusion of some form of editing.

2. Data entry modes

Basically, five types of data entry occur:

- electronic data interchange (EDI). This new instrument has been dealt with in a separate chapter. It is typical for this method that at least part of the editing can be done by the respondent;
- scanning is another hi-tech alternative. Relevant parts of incoming paper forms are translated electronically into bit-mapped information;
- present-day optical character recognition (OCR) systems allow simple edit checks, like valid values and value ranges. The method is particularly suited to large data collections. Obviously, readability is the crucial factor. One should take into account that the method does not enable systematic control on the readability of the data reported. Numbers are more easily readable than plain text, but empty cells are not noticed. Modern OCR packages use a dictionary when recognising words, but every combination of figures results in a valid number. Hand-written material is much more difficult to recognise than typed in data. With OCR the main caution is that it requires very accurate questionnaire layout and printing standards to ensure that the answers can be read by the sensors correctly;
- 'heads-up' data entry means that a computer program gives instantaneous feed-back to the data typist entering the form data (UN, 1994);
- 'heads-down' data entry means that data are typed in at high speed and without feedback from the machine. This method may still apply when the quality of incoming data is high. At present heads-down data entry will frequently be 'out-boarded'.

In general, data entry asks for very specific hardware and software. Some valid value checks and range checks can be applied in these systems, so simple typing errors can be detected. The checks can only be soft, as the only thing the typist can do is compare the entered data with the data on the form.

On pc's database programs and dedicated programs like Blaise can be used for both heads-down and heads-up data entry. A spreadsheet program is less suited for data entry, because of the

problems with protecting the typist and the data against the consequences of pressing the wrong keys. Only when the typist is familiar with the spreadsheet program this is a minor problem.

For heads-up data entry a dedicated statistical package like Blaise is to be preferred. With this tool, establishing a complete set of checking and routing rules is child's-play. Different types of checks can be applied, like range or valid values checks, relational checks and even imputation of values. Technical details like record format, screen layout, processing of key-presses and file management are dealt with quite automatically. Meta-data and export to other software packages can also be managed in an easy way.

A database program offers features approaching those of a dedicated tool. For more elaborate checks scripts have to be made up. This will require more effort than doing the same thing with a dedicated tool. Both database programs and dedicated tools offer multi-user data entry and data matching with data from other sources. In common or third generation programming languages (PASCAL, BASIC, C etc.) the technical details have to be programmed repeatedly for each application. Therefore their usage for simple statistical systems is not recommended.

Even when sophisticated libraries like TURBO VISION can be used, the time needed for developing and testing a system will be a multiple of the time needed to apply a database program or a dedicated tool.

Which of the methods discussed applies, eventually depends on labour resources, available capital and technological know-how. In any case the choice of the method to use should be made in the context of the design of the editing process. Indeed, many aspects of entering are relevant for editing, as will be shown in the next chapter.

3. Timing

When completed forms return to the NSI, the first thing to do is check whether they are (almost) blank. Unusable forms can be considered as non response or can be scheduled for follow-up, depending on the importance of the observation.

It is obvious that one should not wait with entering the data until the entire collection process is completed. Firstly, for most surveys one will wish to release provisional data. This asks for a permanent monitoring of response rates - weighted according to the share of respondents in the total - as well as for ongoing judgement of data estimations in order to determine when the data are reliable enough to issue. Secondly, follow-up actions towards respondents reporting implausible data should be undertaken as soon as possible after return of the form.

4. Editing during data entry

Although there may be good reasons to incorporate certain editing operations during data entry, it is also useful to have the disposal of a data set of raw data as supplied by the respondent. Such a data set enables the statistician to carry out systematic error analysis, which might be interesting for testing of clearness of questionnaires. Furthermore, by saving the original data the value added of editing operations can be determined. Thirdly, during subsequent stages of the processing, discussion might arise as to the correctness of certain edits. This holds in particular when consistency checks with data from other surveys reveal differences between edited data.

In this handbook we will assume that data entry does not comprise editing corrections. Therefore the data file resulting from the collection stage, can be said to be a true duplicate of the set of questionnaires, as completed by respondents.

PART D PROCESSING AND ANALYSIS

Starting point for the processing stage is the information as collected from respondents. Processing and analysis can be said to comprise all operations, applying for "promotion" of these data to the level of the intended statistical output as specified in PART A and as embodied in the table outline.

For various reasons, the act of processing comprises more than just aggregating questionnaire items:

1. part of respondents will make errors while filling in the questionnaire. The same holds for typists when entering the data;
2. both at micro (a) and aggregated (b) level there will inevitably show inconsistencies with related items as obtained from other surveys;
3. some respondents will only partly complete the questionnaire;
4. to comfort respondents, not all of the questionnaire items are a perfect representation of the output concepts envisaged;
5. only part of the elements of the frame population have been surveyed;
6. there will inevitably be non-response;
7. the sampling frame from which the sample was taken, is not a perfect representation of the survey population;
8. certain output data require analysis, like seasonally adjusted indices.

Processing comprises a range of operations, subsequently settling with all of these deficiencies. The following steps apply:

- After data entry, errors (1) and inconsistencies (2a) are detected and corrected during *editing*.
- Subsequently, item non response (3) as well as gaps between questionnaire concepts and output concepts (4a) are dealt with by *imputation*.
- Next, the resulting set of clean and complete micro data serves as the basis for *weighting* (5) and *reweighting* (6). During this stage, also frame errors (7) are accounted for.
- The aggregated data are then confronted with related data from other sources and *integrated* (4b).
- Finally, where appropriate, statistical compilations and *analysis* are carried out, resulting in final tabulations (8).

Each of the steps mentioned accounts for a chapter in this part. It should be noted that the distinction between the steps is not always as sharp as presented here. This goes in particular for the boundaries between editing and imputation and between imputation and (re)weighting.

I. DATA EDITING

Author: Frank van de Pol

1. Introduction

"Put simply, editing is the examination of data for the purpose of error detection" (ABS, 1993). Like any survey respondent, a business statistics respondent is prone to make errors while completing a statistical questionnaire. Only part of these errors can be traced by the statistical agency collecting the data; many errors can and will not be traced. Therefore even exhaustive data editing will never result in an error-free data file. This is not something to really worry about, as long as we manage to trace the important, influential errors. Fortunately, there are methods available to do so.

Before turning to these methods, we will first restate the diverse modes of data entry. Next we give a classification of edit rules. Finally editing methods will be discussed, with special attention for automated editing and selective editing.

2. Setting up a list of edit rules

In general it is easier to describe what is allowed ('checking rules') than stating what cannot be allowed ('rejection rules'; Ferguson: UN, 1994). We will distinguish checks and edits by the moment they occur in data processing and by type. The following terms will be introduced:

- completeness checks
- routing checks
- data validation: valid values checks, range checks
- relational checks, arithmetic checks, ratio edits
- deterministic versus stochastic edit rules

Data editing takes place during or after data entry. *Routing checks* test whether all questions which should have been answered in fact do have been answered. *Data validation* checks test whether answers are permissible. In business surveys the *valid value range* often has to be very wide because of varying firm size. In that case *relational checks* are a powerful editing tool. Many checks take the form of a ratio between two variables, which should be within specific bounds. Another type of relational check is the *arithmetic check*, for instance specifying that a sum of variables should equal a total.

Some edit rules are *stochastic* in the sense that they are just a warning that something might be wrong (soft errors). Exhaustive editing bears the risk of *over-editing*, leading to change of data looking suspect, but nevertheless not in error. Other edit rules are *deterministic*: violation of such a rule points with probability 1 to an error.

3. Organisation of the editing process

Not all editing strategies practised are efficient. We distinguish five alternatives, some of which may be combined to form an optimal strategy:

- paper and pencil
- iteration of data entry and error lists
- computer-assisted data entry and editing
- automated editing
- selective editing

When paper-and-pencil editing is abandoned, and iterative data entry with generation of error lists has been found to be too time-consuming, one often turns to computer-assisted data entry and editing. Then error correction (of all records) can take place during 'heads up' data entry. A drastic solution for editing was developed in Canada and the US. In 1976 Fellegi and Holt published an influential paper on fully *automated editing*. In official statistics its theory was programmed and put into practice in Canada (CanEdit, GEIS), US (Speer), Hungary (Aero), Spain (Dia, Lince) and Italy (Daisy). These programs localise fields probably containing an error, and impute some prediction of the true value. Subsection 4 gives more details. Granquist and others (UN, 1994) explored methods which do editing more *selectively*, developing methods which correct the most influential errors only. Subsection 5 discusses several methods.

When an error check fails and a field has been found to be erroneous, one may simply discard the form (and adapt weights), call back respondents, let clerks make arbitrary 'corrections' (according to general guidelines) or apply automatic imputation, which may be deterministic or stochastic. The latter is preferred, as we will see in the next subsection, in which the issue of automated editing is further elaborated.

4. Automated editing

In the process of editing three stages can be discerned:

1. detection of errors or inconsistencies,
2. in case of an inconsistency: determining which field is in error,
3. correction or imputation.

One may use a fixed *deterministic* rule to determine which field is in error, but in most cases a stochastic alternative is preferable, like the one proposed by Fellegi and Holt. In its simplest form this method states that, if variable A is related to three other variables, B, C and D, an erroneous A-field may generate three inconsistencies, that is with B, C and D (assuming that relational checks A-B, A-C and A-D have been defined). From this observation a natural criterion for determining which field is in error, seems to be the number of inconsistencies a field is involved in.

Example

Suppose a production survey has observed data on revenues (gross output), costs (materials, wages, indirect taxes) and gross operating surplus (GOS). Three relational checks have been defined:

1. revenues = costs + GOS,
2. (GOS / revenues) ≤ 0.4
3. (costs / revenues) ≥ 0.5

The following record shows up:

revenues = 200
costs = 125
GOS = 100

This record violates check 1 because $200 \neq 125 + 100$. Moreover, check 2 is not satisfied because

$$GOS / revenues = 100 / 200 = 0.5$$

which exceeds 0.4.

Both the values for revenues and GOS are suspect, because they occur in two violated checks, whereas costs occurs in one violated check only. Let us suppose that revenues is in error. In that case revenues should be

$$125 + 100 = 225$$

However, GOS remains too high:

$$GOS / revenues = 100 / 225 = 0.444$$

which exceeds 0.4 (check 2). The only remaining solution is to adapt GOS:

$$GOS = revenues - costs = 75.$$

This solution is parsimonious and also satisfies all edits. Note that the deterministic rule ‘always adapt the total when a summation-check is violated’ would not work with this record.

The example shows that the Fellegi-Holt methodology (1976) needs to be more advanced than just counting the number of inconsistencies a field is involved in. To reduce dependence of decisions on the number of checks defined, it involves an analysis of edit checks, removing logically superfluous checks and adding implied checks. Moreover, one should process records as a whole, not field by field, to avoid the introduction of new inconsistencies. The smallest possible set of fields is localised, by means of which a record can be made consistent with all checks. A disadvantage of the method is that the localisation of this smallest set can take up a large amount of computer time.

The Fellegi-Holt methodology often applies ‘*hot-deck*’ imputation. Hot-deck imputation of a failing ‘receptor’ record looks for a similar ‘donor’ record not violating any error check, in order to copy missing fields from. When the definition of ‘similar’ is very demanding, requiring identical scores in many fields, such a record may not be found, and the system will often take refuge to some simple default imputation. Recently in Canada Bankier proposed a much faster

new imputation methodology (NIM) which takes the availability of donor records into account when looking for the set of fields to be corrected.

Example (continued)

In the previous example editing was done using check 1. Now suppose that revenues are not observed, only costs and GOS:

costs = 125, GOS = 100.

In this case another method applies, e.g. , the hot-deck method. This means the searching of a donor record with almost equal costs. The NIM method, however, will look for a record which is similar to the erroneous record, also with respect to the erroneous fields. Two candidate donors are found,

1. costs = 150, GOS = 60
2. costs = 150, GOS = 65

which would produce either of the following corrected records:

1. costs = 125, GOS = 60
2. costs = 125, GOS = 65

Corrected record 2 is chosen, because it corresponds better to the original record. In fact a distance measure is used, which also takes into account the distance between the donor record and the corrected record.

Note that the NIM methodology will generally lead to poor edits when a large error has to be repaired. If the erroneous record would be

costs = 125, GOS = 75000

then NIM may come up with a corrected record like

costs = 30000, GOS = 10000,

which is undesirable if the true record was

costs = 125, GOS = 75.

For numerical data Chernikova's algorithm is presently a good way to implement the Fellegi-Holt methodology. It can handle checks which are related by the logical 'and' operator, but not checks which are related by the 'or' operator. It is the basis of GEIS, which was developed by Statistics Canada. Statistics Netherlands presently develops CHERRYPI, which, in contrast to GEIS, will also be able to handle sum-restrictions, like A=B+C (De Waal, 1996).

5. Selective editing

Selective editing comes down to detection of outliers. It can take place during data entry or when most data have been collected (*macro editing* / *graphical editing*). Editing during data entry (*input editing*) has the advantage of timeliness, because it can start as soon as the first form comes in, but input editing should not be too extensive because it is costly, being record-oriented.

To reduce cost one must be selective, for instance by splitting forms into a *critical stream* and a *non-critical stream*, leaving the latter to automatic procedures. Instead of a simple splitting criterion like firm size, a '*score function*' or '*risk index*' may be constructed.

Hidiroglou and Berthelot (1986) focused on a survey with one important numerical variable and proposed to use a *score function* which takes the ratio of a unit's current value to its previous value as basis for locating erroneous fields. Also the share of that unit in the publication total is taken into account. Van de Pol and Molenaar (1996) stipulated that a large distance of an observed ratio to the median ratio (for firms of a specific size within a specific branch of industry) may be considered as a crude estimate of the size of the error. One may construct a *risk-index*, built up of size of the error and share of the firm, the latter being inclusion probability, non response probability and contribution to total. Furthermore, these ideas were extended to surveys with several target variables.

Example (continued)

Let us return to the first example. Check 1 can be considered as a ratio:

revenues / (costs+GOS)

which should be 1 if the record is OK. Suppose that for the second ratio

GOS / revenues

we expect a value of about 0.1 (last year's median) and hence for the third ratio, we expect about 0.9. A risk index could be computed by dividing each ratio by the median, or dividing the median by the ratio, whatever gives the highest result. This gives the following risk-components:

1. [(costs+GOS) / revenues] / 1 = 225/200 = 1.125
2. [GOS/revenues] / 0.1 = [100/200] / 0.1 = 5
3. 0.9 / [costs/revenues] = 1.44

Before summing them, these risk-components are raised to a power according to the weight attached to the corresponding check. Check 1 is raised to power 5, which gives 1.8; the other ones remain as they are.

Next, the deviations from one of those components that correspond to a violated check are summed to give the unweighted risk index, which is

(1.8-1) + (5-1) = 4.8

for the present record. Note that the unweighted risk index will be 0 for a record that satisfies all edit checks.

Next, the index is weighted, that is multiplied, by a robust estimate of the number of mandays worked by this firm and the number of records in the population it stands for, and it is divided by the number of mandays in the population for the present publication cell. The present firm is a self employed business, which has worked 180 mandays and stands for 20 businesses, whereas 100,000 mandays have been worked by all businesses in this publication cell's population. This gives the weighted risk index

RI = 20 180 4.8 / 100,000 = 4.5

Finally, a standardised OK-index between 0 and 100 is computed by

OK = 100 {1 - RI/[median(RI)+RI]}

With a median weighted risk index of 0.5 this record gets a low OK-index of

100 {1 – 4.5/[0.5+4.5]} = 10

which means it should be edited.

Instead of a risk-index one may use the *Mahalanobis distance* (Little and Smith, 1987). Here over-editing is avoided by inspecting the distribution of Mahalanobis distances, which is known to be chi-squared for multivariate normally distributed data. 'Clean' data should exhibit this distribution (after transformations if necessary) to apply this stopping rule. Alternatively, one may do a pre-study using exhaustively edited data to assess above which risk-index value a record should be in the critical stream. (This also supports the statistician with the composition of the index.) It is even more simple to prioritise record editing according to the height of the

index and stop when publication figures are no longer affected (*top-down* editing; UN, 1994). A prerequisite of the latter method is that the number of publication figures is small enough for visual inspection.

Example

In the previous example one might edit all records with the OK-index below a fixed threshold, say 50. In case of top-down editing this threshold is determined during the process of editing. Publication figures are computed after editing limited chunks of data, say every 10% of the records. At these occasions tables or graphs are made that display the changes in these figures due to editing. When publication figures are stable, editing stops.

Macro-editing or *aggregate editing* is *output editing*. It systematises what every statistical agency does before publication: verify whether publication figures look plausible. To do this one may compare totals in publication cells with the same figures at time point *t*-1. *Graphical* software can be a powerful instrument to show a deviation from the trend. For macro-editing to be efficient it is also important that the software enables to zoom in from the aggregates to the micro data and that changes on the micro-level can easily be made. The gain in efficiency is that errors in non-suspicious aggregates are not edited anymore.

Example

Foreign trade figures are hard to edit. Checking the millions of records that come in each month on a one-by-one basis is a mission impossible. With macro-editing one can select severe errors for editing and skip small errors. From a macro-editing viewpoint trade flows are suspect when they deviate strongly from their historical course. Records that contribute to a large extend to suspect publication figures are checked first. Records that contribute to publication figures that look normal are not checked anymore.

This major improvement is recently developed for the foreign trade statistics of the Netherlands and Canada. Of course, specific formulas have to be devised to predict current trade figures from history and to prioritise editing records related to suspect publication figures.

Finally, it should be noted that selective editing is not completely without risks. Care should be taken that it does not cause too much bias. Bias will occur if for instance only large positive deviations from the expected value are corrected and large numbers of negative deviations (zeroes) are ignored. Also false stability, due to firms who return exactly the same answers at every occasion, can damage the validity of publication figures.

6. Checking external consistency

Thus far we implicitly focused on consistency checks between items from one and the same questionnaire. However, also checks with data from other surveys may apply. This might even be a reason for including redundant items in different questionnaires. “Number of employees” is

a notorious example. External consistency checks, first at the macro (aggregate) level and where necessary also at the micro level are an important means to reduce problems during the integration stage. The applicability of this type of checks heavily depends on the degree of co-ordination among the surveys compared. Once again, here is an advantage of complying as much as possible with the concepts and standards of the coherent framework when setting up the survey.

References

Australian Bureau of Statistics, 1993, Data Editing - an ABS manual.

Hidiroglou, M. A., and J.-M. Berthelot, 1986, Statistical editing and imputation for periodic business surveys. Survey Methodology 12, pp. 73-83.

Little, R.J.A., and Ph.J. Smith, 1987, Editing and imputation for quantitative survey data. Journal of the American Statistical Association 82, pp. 58-68.

UN, 1994, Statistical data editing methods and techniques, vol. 1. (UN Economic Commission for Europe, Statistical Division, Geneva)

Van de Pol, F.J.R., 1994, Selective editing in the Netherlands Annual Construction Survey. paper presented for the work session on statistical data editing in Cork, Ireland (UN Economic Commission for Europe, Statistical Division, Geneva)

II. IMPUTATION

Author: Peter Verboon

1. Introduction

In most surveys one has to face the problem of missing data. Two types of missing data are usually distinguished: *unit* non response and *item* non response. Imputation applies for the latter, while unit non response is dealt with by reweighting. Methods to employ in case of unit non response will be discussed in the next chapter.

There is a third manifestation of missing data, caused by the gap between output variables desired and questionnaire items surveyed. Unlike non response, it is deliberately decided on during the survey design stage. Therefore we will speak of *intentional missing data*.

2. Item non response

Item non response or partial non response occurs when the sampled units have not answered all relevant questions, but did respond to part of them. (If a respondent reported on *all* questions, inconsistencies between some of the answers may occur or an answer may be logically incorrect. This problem relates to editing, and has been dealt with in the previous chapter.

Little and Rubin (1987) distinguish three different types of item non response. In the first type the missing values are completely at random. The second type does not depend on the value of the variable, but on the values of some other variable(s). The third type depends on the value of the variable on which it is missing, e.g. high scores are more likely to be missing than small ones.

Example

Suppose a form containing an item about pollution tax. If enterprises do not respond to this item, because they erroneously assumed the question was not meant for them, the non response is called *completely at random*. Indeed, whether or not the question was answered has no relation with the level of the tax nor with any other item.

Now suppose some of the businesses use to pay pollution tax for a particular kind of chemical they expose to the air. Recently this chemical has proved to be rather dangerous for the environment. Such respondents may be more likely to "forget" the pollution tax question than other respondents. In this case the non response depends on another variable, i.e. type of pollution. Now we might say that, *given the type of pollution a business is taxed for*, the non response is completely at random.

The non response is not at random when businesses with high tax levels are more likely not to respond than firms with low levels of tax.

Before applying any method to handle the missing values, it is important to find out what type of non response one is dealing with. In Kim and Curry (1977) and Frane (1977) procedures are given to test whether the non response pattern is random. The decision which method to use can be based on the type of non response.

3. Methods for handling item non response

Two general strategies apply with respect to the way to deal with item non response.

The first strategy *ignores* the missing values while their treatment is deferred to the analysis stage. The most simple way is to ignore all forms with missing values and confine to analysis of the fully completed forms. This method is called the *complete case analysis.* An alternative is the *available case* method, which uses all available information for univariate and bivariate statistics. More elaborate methods directly analyse the incomplete data by specifying a model. Methods which directly analyse the incomplete data are based on distribution assumptions regarding the variables.

In the second strategy estimates for the missing data are sought and with these new values the data matrix is made complete. This is called imputation. On the imputed data matrix standard analysis techniques can be applied.

Direct methods

A *complete case* analysis can usually be obtained in standard computer packages by choosing the option "listwise deletion". This method is easy to understand and readily applicable, but by deleting all cases with one or more missings the sample size can become very small.

The *available case* method uses more information than the complete case approach. Univariate statistics, like the mean, are computed for all available (non missing) cases for that variable. And bivariate statistics, like the correlation coefficient, are computed for all available pairs. Problems arise when surveys have different sample sizes. Furthermore, the correlation matrix may not be positive definite. For both methods it is implicitly assumed that the missing values are completely random. In the more elaborate approaches a model must be specified. This model contains the parameters of interest. Furthermore, an assumption about the distribution of the variables must be made. Usually the data are assumed to follow a multivariate normal distribution. Based on such a model a likelihood function can be defined, which is maximised over the parameters. The well-known EM-algorithm has been developed in this context.

The E-M algorithm in a nutshell

E and M stand for *Expectation* and *Maximization*, which refer to the two steps of the procedure. The procedure alternates between these two steps until it stabilises. In the Maximization step the parameters of the model (e.g. means and covariances) are estimated based on the observed sample. In the expectation step updates for the missing values are computed based on the model with the parameters from the previous step. Repeating these two steps a sufficient number of times will always yield a stable solution for the parameters of the model and for the missing values.

Imputation methods

There are a variety of imputation methods, ranging from very simple and intuitive to rather complicated statistical procedures. The most important methods will be mentioned briefly. For a more extensive discussion we refer to Sande (1982) and Allen (1990).

- Subjective treatment: impute on the basis of values which appear reasonable. For example, one might deduce the labour costs if the number of employees is known.
- Mean/modus imputation: impute the mean of a variable or the modus (for categorical data). An improvement may be to impute the median in order to eliminate the effect of outliers.

- Post stratification: divide the sample into strata and then impute stratum mean/ modus/median.
- Cold deck imputation: find reasonable estimates for the missings in another data set, for instance a previous measurement (historic data) or another source close to the non responding one (proxy data).
- Hot deck imputation: find a donor case in the data set. First make homogeneous imputation classes and then select a donor at random.

Example

The variable "profit" can be imputed by first listing groups of enterprises which produce similar products. Next the profit of one enterprise, selected at random, can be used for a respondent with missing profit data, provided both enterprises belong to the same group.

- Nearest neighbour imputation: select a donor which is close to the non respondent according to some distance criterion. Or select one from the nearest group.
- Regression imputation: define predictor variables and estimate the missing value. Add random error to the prediction to compensate for underestimation of the variance. The predictor variables can also be chosen in an optimal way.
- Predictive mean matching: combination of the regression and nearest neighbour approach.
- Multiple imputation: impute several values for the missing variable in order to obtain better estimates for variances and covariances.
- Simultaneous imputation: impute different values from the same record to ensure consistency.

Example

Suppose a questionnaire for which both *turnover* and *profits* are missing. Now one can take both values from one and the same donor respondent to ensure consistency between the two related items.

4. Intentional missing data

This type of missing data occurs when during the questionnaire design stage it was decided to refrain from explicitly surveying certain target variables as chosen and defined during the output design stage. Such a gap may have been created deliberately in order to better fit in with respondents accounting systems, or because certain specifications are not worthwhile to bother smaller respondents for.

Example

For the compilation of data on "fixed capital formation", a surveyor wishes to measure the purchase value of new fixed assets put into use. However, enterprises that lease the assets acquired, will not be able to supply the purchase value. Therefore, the questionnaire mentions "lease amounts paid", and the NSI imputes the purchase value by means of certain keys.

Imputation also applies when the respondent is not able or willing to report data at the level of the statistical unit envisaged. In that case the data supplied for the *reporting* unit will have to be converted to the *statistical* unit applying.

Example

For the KAU-based monthly survey on turnover, an enterprise is split into two KAU's. However, the respondent is willing to report for the enterprise as a whole only. The NSI imputes the turnover data for the KAUs by allocating the total figure according to one of the methods discussed above.

Intentional missing data will become an increasing phenomenon in business statistics, as the impact of policies towards reducing respondent burden is growing. In principle, the imputation methods described in this chapter apply both for unintentional and intentional missing data.

5. Discussion

Which method applies best depends on the objectives of the analysis and on the type of non response. If the objective is to estimate population means and population totals, there is no interest in the individual scores. Imputation is less important here, and a direct analysis may be a good choice, assuming that the distribution assumptions are reasonable. Sometimes the main interest is in individual values, while population parameters are of less importance. Here some imputation method is clearly appropriate. The main conclusion is that no method is superior to the others in all circumstances.

6. Summary

Imputation applies in the first place when questionnaires are not filled in completely. Furthermore, imputation techniques can be used to bridge the gap between *questionnaire* items, reported at the level of *reporting* units, and *target* variables as applying for *statistical* units. This transformation means a fundamental change in the nature of the data in the data flow, which will from now on be *output* oriented.

References

Australian Bureau of Statistics, 1993, Data Editing - an ABS manual.

Hidiroglou, M. A., and J.-M. Berthelot, 1986, Statistical editing and imputation for periodic business surveys. Survey Methodology 12, pp. 73-83.

Little, R.J.A., and Ph.J. Smith, 1987, Editing and imputation for quantitative survey data. Journal of the American Statistical Association 82, pp. 58-68.

UN, 1994, Statistical data editing methods and techniques, vol. 1. (UN Economic Commission for Europe, Statistical Division, Geneva)

Van de Pol, F.J.R., 1994, Selective editing in the Netherlands Annual Construction Survey. paper presented for the work session on statistical data editing in Cork, Ireland (UN Economic Commission for Europe, Statistical Division, Geneva)

III. WEIGHTING AND REWEIGHTING

Authors: Elly Koeijers and Karin Hilbink

1. Introduction

Samples result in information for only part of the target population. This chapter discusses methods to expand this information to the level of the target population. It is common practice for statistical offices to attach *weights* to the elements in a sample. Throughout this paragraph we assume that the parameters of interest are population totals. We assume that the target population consists of some type of statistical unit and that the sample has been drawn according to a chosen sampling design.

Objectives of weighting are:

1. expand the sample to the population.
2. cope with missing observations.
3. increase precision by utilisation of auxiliary information.
4. achieve consistency with data from other sources.

We distinguish between weighting and reweighting. Weighting, i.e. the attribution of weights to sampled units, can in principle take place *before* data collection, provided the sampling design is not too complex. Reweighting always applies *after* data collection.

2. Weighting

Expansion of the sample to the population is a process in which first order inclusion probabilities play a key role. The first order inclusion probability of a population element is the chance that this element is included in the sample. If the population for the target in question consists of N enterprises, N first order inclusion probabilities are associated with the chosen sampling design, say $\pi_1,..., \pi_k,..., \pi_N$.

The inverse of π_k is called the *inclusion weight* of element k, denoted $a_k = 1/\pi_k$. The term *weighting* usually refers to the inclusion weights. In a business survey where the population is stratified before sampling and where the samples within the strata are drawn according to simple random sampling, the inclusion weights for the elements in stratum h are $a_k = N_h/n_h$, where N_h is the population size and n_h is the sample size in stratum h.

Example

Suppose the population of computer services industry is divided in the three strata *small*, *medium* and *large*. A possible allocation of the sample and according inclusion probabilities is given in the following table.

Stratum (h)	*small*	*medium*	*large*

population (N_h)	1000	300	50
sample (n_n)	250	150	50
inclusion probability (π_h)	¼	½	1
inclusion weight ($1/\pi_h$)	4	2	1

The inclusion weights are the inverses of the inclusion probabilities.

3. The π-estimator

The well known π-estimator (or Horvitz-Thompson estimator) is directly based on the inclusion weights. Its values are computed by summation over the observations multiplied by their inclusion weights. The specific form of the estimator is determined by the first order inclusion probabilities originating from the chosen sampling design. The π-estimator utilises auxiliary information from the design stage only. In other words, the inclusion weights do not depend on the values of the variables observed in the data collection process of the survey. However, in case of non-response it is necessary to adjust the inclusion weights.

Example (continued

Suppose we want to measure total turnover of the population in the strata, whereas the turnover (in million dollars) measured in the three strata amounts to:

Stratum (h)	*small*	*medium*	*large*
turnover	4	50	1000
inclusion weight	4	2	1

The total turnover in the population is estimated as:

4x4 + 2x50 + 1 x 1000 = $ 1.116 million.

When some of the sampled businesses do not respond, the inclusion weights have to be adjusted for non response. This is illustrated in the following table.

Stratum (h)	*small*	*medium*	*large*
population (N_h)	1000	300	50
sample (n_n)	250	150	50
inclusion probability (π_h)	¼	½	1
inclusion weight ($1/\pi_h$)	4	2	1
response	200	120	40
response probability	4/5	4/5	4/5
adjusted weight	5	5/2	5/4

The adjusted weights are the inverse of the product of the inclusion probability and the response probability.

Now suppose the observed totals of turnover in the three strata are respectively: 3, 3; 46 and 810. The estimated total is:

5*3,3+2,5*46+1,25*810=1084,2 million

Under the condition that no observations are missing and no errors are made during the data processing, the estimator is unbiased for the totals in question. It is assumed that every element of the population has a positive chance to appear in the sample.

In some cases statistical offices deliberately exclude certain elements belonging to the target population from sampling and guess their contribution to the desired total, either because they are too small or volatile to maintain a frame for, or it is impossible to collect data for them. Great care must be taken in these cases.

Example

We follow the four "layers" of a population, as categorised in A-III-2.

Assume that users want turnover data on the computer service industry in 1996, including "hobbyists", rendering incidentally services to acquaintances and not being registered as businesses. Nevertheless, the NSI decides to exclude the latter category from the data.

1. The *ideal target population* comprises the whole industry, including hobbyists;
2. The *target population* excludes the hobbyists, and is therefore a subset of 1;
3. The *business register* does not comprise self employed businesses, and therefore does not cover the whole target population;
4. The *frame population* is a subset of the business register, and therefore does not comprise self employed computer service businesses. Moreover, businesses which died in the course of 1996 are deliberately excluded from the frame population, because it is impossible to collect data from them.
5. The *sample* is a subset of the frame population.

The sample is expanded to the frame population. Further expansion of the latter to the target population demands additional "guestimates" for self employed businesses and recently died businesses.

4. The variance of the π-estimator

First and second order inclusion probabilities are important for the calculation of the variance of the π-estimator. Second order inclusion probabilities relate to pairs of elements in the population. A second order inclusion probability is the chance that two elements are simultaneously included in the sample. The variance of the estimator also depends on the complete set of population values of the target variable under consideration. This is the reason why the variance can seldom be calculated and therefore has to be estimated in almost all practical situations.

For more details on the estimator and variance estimation we refer to Hedayat and Sinha (1991) and Särndal, Swensson and Wretman (1992).

The variance of the estimator becomes exactly zero if the sampling design is of fixed size and the design yields first order inclusion probabilities proportional to the values of the target variable. This interesting property can serve as a tool to construct a sampling design. The values of the target variable are, however, unknown (otherwise the survey would not be necessary). Therefore a survey statistician looks for auxiliary variables that are correlated with the target variables in order to come as close as possible to the ideal situation of zero variance.

Example

If we are interested in the total turnover of computer services industry, the estimator could have a zero variance when we choose the inclusion probabilities proportional to the values of the turnover. However, the values of the turnover are not known. An auxiliary variable that is correlated with the turnover of a business is the size of the business. To come close to the ideal situation of zero variance the inclusion probabilities are chosen proportional to the number of employees.

The largest businesses of the population for a survey are usually included with certainty and their inclusion weight is set equal to 1. The consequence is that they do not contribute to the variance. This is actually an attempt to attain the above mentioned ideal situation.

5. Domain estimation

The usual technique to handle the domain estimation problem (see B-IV.10) involves the definition of a new, artificial variable which is given the value of the variable of interest for elements inside the domain and zero otherwise. The standard approach using the π-estimator can now be applied. In principle the estimate for a specific domain total can be computed over the entire sample, but in practice the zero values of the new variable make that the summation can be restricted to the part of the sample that happens to fall in the domain.

Example (continued)

Suppose the survey has originally been designed at a national level, i.e. measuring total turnover for the country as a whole.

Now suppose we want to derive data on turnover in a certain geographic area. We can treat the area's as *domains* and make estimations for the values in the domains. As explained in the text we define an artificial variable which is given the value of the turnover for elements inside the area and zero otherwise. The estimate for a specific area can be computed over the entire sample. Instead of the original turnover we use the artificial variable in computing the turnover in the region of interest.

The artificial zero values can have a substantial impact on the variance of the π-estimator. The smaller a domain, the less elements of the total sample tend to fall in this domain. Therefore it is important that the sample size is large enough to produce domain estimates with sufficient accuracy. In cases where the domain size is known or other auxiliary information is available alternatives to the π-estimator are commonly preferred.

The method of domain estimation can be applied in surveys for which the frame contains elements that do not belong to the target population. When the frame has overcoverage but no undercoverage, the target population is a domain of the frame population. If we assume that it is possible to recognise the nonpopulation elements when they appear in the sample, then the π-estimator is unbiased for the estimation of population totals.

Example

When a business register contains enterprises having recently gone out of business, but which are still included in the register, the non-population firms should not be immediately removed from a sample, but given zero values on the variables of interest. They should contribute to variance estimates.

6. Reweighting

From the preceding paragraphs it is clear that the inclusion weights can be used for the first objective (to expand the sample to the population). The other objectives are commonly attained by adjusting of the inclusion weights. The adjustment procedure is called reweighting. This is done on the basis of auxiliary information available from administrative registers or from the business register at the moment the estimation takes place.

Example

Suppose again we are interested in regional turnover. From another source the number of businesses in the regions are known. We want to make a publication in which both the numbers of businesses and total turnover in a region are shown. To attune these numbers in the publication we can reweight the turnover to the number of businesses in a region.

Considerable literature exists on the utilisation of auxiliary information in the estimation stage. Särndal, Swensson, and Wretman (1992) give an exhaustive overview of the methodology. Widely used techniques are regression estimators, ratio estimators, ranking ratio estimators, and post stratification.

Non response in a business survey deserves more attention when it occurs for large firms than for smaller ones. Which response rates are acceptable is hard to say, but it is advisable to keep the response rate high among the larger firms, because these usually have considerable impact on the accuracy of most outcomes.

Often businesses are drawn according to a design based on somewhat outdated information, for example according to size class and economic activity from a previous period. In such cases it might be desirable to apply reweighting. Once the observations are ready for the estimation step, more up to date information on size class and economic activity might be present in the frame. These values can be used to adjust the inclusion weights. This is also a way to ensure consistency between different surveys with the same target population, even when the samples were selected from different versions of the business register.

Example

At the time the sample was drawn the size class of the Enterprise was derived from data of the year before (t-1). After processing the results, data of the size class is available for the year t. For every Enterprise it is known in which one of the adjusted size classes h' it is placed. To adjust the weighting factor $w_k = 1/\pi_k$ an extra factor $c_{h'}$ is used for every business. The value of the correction factor $c_{h'}$, follows from $\sum_{k \in h'} \frac{c_{h'}}{\pi_k} = N_{h'}$, where $N_{h'}$ is the number of businesses in size class h'. The expression for the correction weight is: $c_{h'} = \frac{N_{h'}}{\sum_{k \in h'} \frac{1}{\pi_k}}$. The adjusted weight is $w_k = \frac{1}{\pi_k} c_{h'}$.

The issue of consistency has gained renewed attention in two recent articles by Deville, Särndal, and Sautory (1993) and Deville and Särndal (1992). In their approach the inclusion weights are adjusted by means of a set of *calibration equations*. To keep the adjusted weights as close as possible to the inclusion weights a *distance function* is chosen. Several distance functions are proposed, some of which leading to well known results. An advantage is that this technique allows to specify lower and upper limits for the adjusted weights in order to avoid negative or extremely large weights.

7. Frame errors and estimation

The possible existence of frame errors complicates the estimation process. In Lessler and Kalsbeek (1992) an extensive treatment is given of different types of frame errors. They distinguish six categories, four of which are relevant in the estimation stage of business surveys. The four categories are:

- Undercoverage (missing units).
- Overcoverage (inclusion of nonpopulation units).
- Duplicate or multiple listings.
- Incorrect auxiliary information (size, activity, misconstruction of units, etc.).

Undercoverage is perhaps the most serious problem, because it can neither be detected from the frame nor from a sample. Population totals will almost always be underestimated (assuming that the variables of interest have positive values), because part of the target population is out of the observation scope. Estimators for totals are therefore negatively biased. Whether other estimators, for example for proportions, are biased as well depends on the characteristics of the missing part.

Errors of the fourth type (incorrect auxiliary information) may lead to overcoverage. For example, businesses with an incorrect activity code, in the survey not detected as such, may be erroneously be treated as if they are population elements. The reverse, i.e. wrongly excluded elements, may also occur. This leads to undercoverage. Whether effects of overcoverage and undercoverage tend to cancel each other out, depends on the variables being measured.

Estimation in the presence of frame errors is complex. In Särndal, Swensson, and Wretman (1992) three special situations are considered:

- The frame has overcoverage, but no other imperfections.
- The frame has duplicate listings, but no other imperfections.
- The frame has undercoverage, but no other imperfections, and an auxiliary total for the entire population is available .

For each of the above mentioned hypothetical situations a solution is proposed. The first situation actually is the domain estimation problem as discussed in the previous section. In practice most surveys suffer from more than one category of frame imperfections. A general solution solving all frame problems simultaneously cannot be given.

References

Australian Bureau of Statistics, 1993, Data Editing - an ABS manual.

Hidiroglou, M. A., and J.-M. Berthelot, 1986, Statistical editing and imputation for periodic business surveys. Survey Methodology 12, pp. 73-83.

Little, R.J.A., and Ph.J. Smith, 1987, Editing and imputation for quantitative survey data. Journal of the American Statistical Association 82, pp. 58-68.

UN, 1994, Statistical data editing methods and techniques, vol. 1. (UN Economic Commission for Europe, Statistical Division, Geneva)

Van de Pol, F.J.R., 1994, Selective editing in the Netherlands Annual Construction Survey. paper presented for the work session on statistical data editing in Cork, Ireland (UN Economic Commission for Europe, Statistical Division, Geneva)

IV. STATISTICAL INTEGRATION

Authors: Ad Willeboordse and Winfried Ypma

1. Introduction

Up till now we described surveys as more or less isolated entities, existing apart from each other. Also, the collection and processing of the data would evolve separately for the different surveys. Consequently, the data generated after weighting and reweighting can be considered as a self-contained statistical result of each particular survey.

We also stressed many times that all individual business statistics actually are to be considered pieces in a jig saw, in the sense that they together provide the picture of the economy as a whole. This is why it is so important that all surveys are conceptually subordinate to one general framework, as discussed in PART A.

Still, even in a situation where *concepts* of variables and classifications are standardised to a large degree, confrontation of the *data* drawn from different surveys will reveal discrepancies and inconsistencies. The act of *integration* can be said to comprise all activities:

- directed towards the creation of consistency in originally inconsistent data,
- in such a way that the data fit to the integration framework, which is for business statistics primarily the System of national Accounts.

The notion of integration is, indeed, closely associated to national accounts practices. This does, however, not mean that the involvement of surveying statisticians is confined to their role as data suppliers. Integration may very well lead to a higher accuracy of the data of basic statistics as they are provided to external users, and integration will in any case enhance coherence with outcomes of other surveys.

2. Causes of differences

Example

Suppose within an NSI there are two different surveys (e.g. labour and production statistics) which generate data on labour cost in computer service industry on 31 December 1995. Almost “by definition”, the outcomes will deviate. Users, being confronted with both figures, will get confused.

There may be a thousand reasons for the figures not to correspond. All of them can be ascribed to either of two major categories, which have to be distinguished clearly:

- conceptual differences, referring to the use and definition of variables, units and classifications;
- operational differences, referring to the observation of concepts, i.e. to methods of collection and processing.

Example (continued)

Suppose:

- the labour survey is based on *local units*, while the production survey is based on *enterprises*
- the labour survey includes the salaries of hired staff, while the production survey does not.

These are examples of conceptual differences: the concepts for *industry* and for *labour cost* are not the same in both surveys. In case of such conceptual differences, it is unjust to say that the survey results are *inconsistent.* Actually, they appear to measure different phenomena. What is wrong here is the terminology, i.e. the use of the same names ("industry", "labour cost") for different "things".

Example (continued)

Suppose the concepts of the two surveys are made equal, so that in both surveys the data are observed according to the same concepts.

Still, the outcomes continue to differ. One of the (many) reasons appears to be that the labour survey is based on a sampling frame as established on December 31, 1995, while the production statistics is based on the frame as it exists on March 31, 1996

This type of differences is of a fundamentally different nature. Now the difference is not in the concept as such, but in its *observation*. In this case the differences can be justly qualified as inconsistencies. There are many more reasons for inconsistencies, among which (differences in) sampling errors and non-sampling errors for the two surveys.

Example (continued)

Suppose National Accounts use the data from the production survey as an input for their supply and use tables, in the understanding that the figures received are raised for black labour cost. The result is that National Accounts data for labour cost in the computer service industry differ from those in basic statistics.

The question is whether this difference is of a conceptual or an operational nature. The answer depends on the definition of labour cost in the basic statistics. If labour cost is defined as "wages and salaries etc. *according to payroll*" of the respondent, the difference is *conceptual*. If the payroll addition is explicitly excluded, the concepts are the same and thus the difference denotes an inconsistency. In the latter case, there is reason to adjust the labour cost figure of the basic statistics, because it leads to a figure that better meets the definition. If adjustment also takes place in the in the first case, this would come down to a change of the concept of labour cost in the basic statistics.

3. The act of statistical integration

In a broad sense statistical integration comprises the set of all measures and activities aiming at cohesion of statistical data. We will now provide an overview of such measures and actions. Because the focus of the Handbook is on surveying we will confine ourselves to measures that fall within the scope of basic statistics.

- attuning of concepts regarding variables, statistical units and classifications among basic statistics.

Example (continued)

In the case of the labour survey and the production survey with their deviating definitions of *labour cost* and *industry*, basically two solutions apply:

1. adjust the *names* of the concepts, so as to stress that both surveys observe different things. This is, of course, not an act of integration. Rather, it is the opposite, i.e. the admission that the two concepts cannot be related;
2. adjust the *definition* of the concepts of one survey to those of the other. This supports integration, in the sense that the two surveys are now conceptually linked.

- attuning of concepts between basic statistics and National Accounts. In this respect, the *supply and use table* as developed in ESA is an important help. It is deliberately designed to bridge the gap between the more theoretical and analytical parts of the National Accounts and the practically oriented basic statistics. Therefore, its concepts are chosen as "realistic" as possible.

Example (continued)

From an integrative point of view, basic statistics and Supply and use tables should use the same definition of labour cost. This might lead to the conclusion that the basic statistics have to include black labour in their definitions. The fact that black labour is not observable in ordinary collections is no reason to abstain from this measure. Indeed, the missing data can be provided by National Accounts, be it at meso/macro level only.

- eliminate duplications.

Example (continued)

After cleaning up the conceptual differences between labour survey and production survey, it seems self-evident to conclude that one of the surveys can skip the labour cost question and derive the data from the other.

This is, indeed, the most rigorous way to get rid of contradictory information, and, as we have seen earlier, it has other advantages as well. In practice, however, following this trivial suggestion appears to be far from simple. Survey managers use to raise all kinds of objections, mostly of an organisational nature. Such objections should, however, in general be subordinate to external considerations, of which response burden (to comfort data suppliers) and coherence (to comfort users) are prevailing;

- Adopt a "one figure" doctrine.
  Assuming that it will never be possible to skip all duplications, and thus taking for granted that different figures for the same items are *produced*, an NSI can take measures in order to prevent these figures from being *published*. This asks either for appointing one survey a prevailing position or for making agreements which require that individual survey managers stick together and agree on which figure to publish.
  A centrally maintained *data warehouse* as described in PART E is a powerful tool in this respect.
- Centralise statistical processes.
  The best known illustration of this measure is the Central Statistical Business Register, which prevents the different survey units of making their own, and consequently deviating, survey frames. It should, however, be noted that the pure existence of such a central register (and the absence of decentralised registers), does not guarantee integration of populations.

Example (continued)

We saw already that the labour survey defined its sampling frame according to the Central Business Register as it was on December 31, 1995, while the production survey took the picture three months later. Although both surveys are based on one central register, their populations will differ.

The example illustrates that central agreements on the *use* of the Business Register apply as well. Another opportunity to centralise and concentrate parts of the statistical process is provided by electronic data collection (EDI). EDI is a natural drive in this direction. It is very unlikely that respondents agree in having their record keeping systems tapped from different points in the NSI at different moments.

- Large Business Unit.
  Large businesses take a considerable share in the outcomes of statistics. At the same time their complex and frequently changing organisation and activity structure makes them liable to observation errors and, in case of deconcentrated surveying, for inconsistent reports. These risks can be reduced by the establishment of a Large Business Unit, which at least takes care for continuous profiling of large and complex businesses, and preferably also for the data collection of the major business statistics.

- Large Business Data Base
  Integration by National Accounts is to a large extent confrontation of aggregated data from different sources, taking place *at the end* of the statistical process. One of the consequences is that already published data from basic statistics have to be reviewed.
  Therefore, the ideal situation would be to undertake the confrontations at the earliest possible moment in the statistical process, preferable at the micro level. Such an operation may be specially worthwhile for the larger businesses. Consistency is established at the very beginning of the process and editing can benefit from it as well.

4. Summary

Statistical integration is not the exclusive area of National Accountants. Surveying statisticians are interested parties as well and they have a responsibility in creating a favourable climate for integration. Both respondents, statisticians and users can benefit from integration. Smart confrontations can produce "new" data that otherwise would have been collected, they enforce the efficiency and quality of data processing and they enhance coherence of published data.

V. ANALYSIS: SEASONAL ADJUSTMENT OF TIME SERIES

Authors: Kees Zeelenberg and George van Leeuwen

1. Introduction

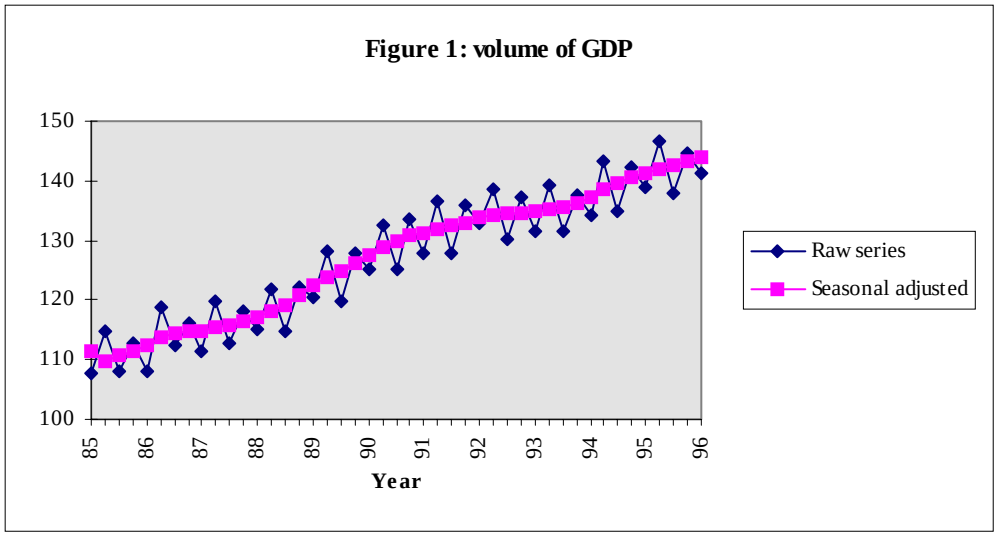
In the context of this manual the stage of analysis refers to statistical operations, applied on the data as they result from the processing stages. There is an almost infinite variety of analyses, depending on the nature and objectives of the survey. A practical handbook like this, is not the place to discuss all possible analytical methods. Therefore, as an illustration, we will confine ourselves to an analytical operation frequently applying on data from business surveys, i.e. seasonal adjustment of time series.

2. The need for adjustment

Many economic time-series show cyclical fluctuations. This is most obvious for series published with a period less than a year, such as stock prices and unemployment. But also series with a period of a year may show cycles; for example inventories often have a cycle of a few years. If the period of the series is less than a year, then fluctuations with a cycle of a year are called *seasonal* fluctuations.

Major causes of such seasonal fluctuations are *calendar effects* (for example, the number of working days in a month), *institutional effects* (for example, because youngsters usually leave school in June, the months June, July and August often show a higher level of unemployment than May and September), and the *weather* (for example, due to bad weather in the winter there is during the winter months an increase in unemployment of construction workers).

If a series shows seasonal fluctuations, one might want to compute a new series that is somehow corrected for these seasonal fluctuations. A major reason why we might be interested in such a *seasonally-adjusted* series is its policy relevance: if for example unemployment in a certain month increases sharply, policy makers need to know whether this increase is, more-or-less, permanent or due to a seasonal factor. Seasonally adjusted data are also a pre-requisite for assessing the state of the business cycle. In figure 1 is shown that a clearer picture of the state of the economy - as measured by the development of the Gross Domestic Product (GDP; here for the Netherlands) - emerges after having seasonally adjusted the data.



3. Adjustment methods

The principle of decomposition

In general, adjustment methods presuppose that a series can be divided into three components:

- the *trend and cycle*: the trend is the general movement of the series over a number of years and the cycle consists of fluctuations longer than a year;
- the *seasonal component*: fluctuations with a period of a year;
- the *irregular component:* fluctuations of a purely accidental nature.

The division may be either additive or multiplicative; in the first case, the seasonal component of the series is independent of the size of the value of the series, and in the second case it is proportional to the size.

Having only observations on the series, the analyst wishes to learn about the three *unobserved* components by using a decomposition method which gives an estimate of the seasonal factors, the trend-cycle and the irregular component. Applying the seasonal factors to the original series yields the seasonal adjusted series: the sum of the trend-cycle and the irregular component.

There are several methods for obtaining a decomposition of the time series into the three unobserved components. These can be summarized into two broad classes: the so-called Census methods and the 'model-based approaches'. The most well-known example of the class of Census methods is the *Census-X-11 variant,* developed by the US Bureau of the Census. Recent variants of the Census method are *X-11-ARIMA* (developed by Statistics Canada) and *X-12-ARIMA* (developed by the US Bureau of the Census). Examples of the 'model-based' approach are the *Structural Time-series Method (STM)* and a method based on 'signal extraction theory', implemented in the program *Signal Extraction in ARIMA Time Series (SEATS).*

'Pre-treatment' of time series

Before a decomposition method is applied to a series, one usually eliminates outliers, structural breaks and causal effects. Examples of causal effects are calendar effects due to e.g. inter-period variation in the length of months or quarters and the patterns of working days. Other causal effects may be captured by estimating regression effects using auxiliary data. For example, above-average effects of very bad winters may be estimated in a regression of the series on a temperature variable. These so-called 'pre-treatments' aim at 'purifying' the irregular component in order to linearize the series and to obtain a smoother seasonal and trend-cycle component. More advanced adjustment methods such as *X-12-ARIMA*, and *TRAMO* - a 'pre-treatment'

program for *SEATS* - give an indication when they suspect a particular observation to be an outlier or a structural break and are also able to estimate calendar and other causal effects using regression techniques. This enables the user to 'pre-adjust' the series before applying the decomposition.

Decomposition methods

The most simple seasonal adjustment takes place by comparing ('prior adjusted') observations with the value of exactly 12 months earlier. However, this is a very crude method: it takes only two observations of the series into account and thereby ignores the information contained in the intervening periods.

Most other adjustment methods work by taking moving averages of the series; such moving averages are called *filters*. The length of the filter determines whether short-term fluctuations show up in the trend-cycle component or in the seasonal component.

Census methods

The basic principle of the Census methods is the application of 'built-in' fixed filters, so the weights for applying the moving-averaging procedure are independent of the series. The decomposition in the Census methods can be summarized as a long iterative sequence of applying fixed filters for the trend-cycle and the seasonal component in turn. As a consequence the seasonal adjusted figure can be seen as the final result of applying one long filter to the data. For observations sufficiently distant from the last observation the filters are symmetric. Because in the *X-11 variant* proper moving averages cannot be computed for the most recent observations, symmetric filters at the end of the series are replaced by asymmetric weights. In the *ARIMA variants* of the Census method one may extrapolate the series, which is usually done by univariate models. This enhances the application of symmetric weights for all available observations.

Model-based methods

A limitation of the Census methods is that they do not take into account the statistical properties of the series. In answer to this problem, 'model-based' methods have been developed. These methods rely on stochastic linear models for the series and the components under analysis. Examples of 'model-based' methods are - as mentioned before - the *STM-method* and *SEATS*. In the *STM method* one starts by specifying directly a model for the trend-cycle, the seasonal and the irregular component and then estimates the parameters of interest. Alternatively one may start by estimating an ARIMA model for the observed series and then derive ARIMA models for the components that are consistent with the model for the observations. This procedure is followed in *SEATS.*

Having estimated the parameters of all models, the so-called 'Wiener-Kolmogorov (WK) filters' for the components can be derived. These filters represent the ('model-based') weights to be used for computing the moving averages for the seasonal and the trend-cycle component. By construction a WK filter adapts itself to the data under consideration and this adaptation avoids the danger of over or underestimating of e.g. the seasonal component inherent in the fixed filter methodology. An other advantage of the 'model-based' methods is that they give statistical results, such as standard errors of the parameters and a fit of the model, so that proper statistical tests can be applied. However, the 'model-based' approaches share the disadvantage of being rather complicated. Because they are based on the application of very sophisticated statistical techniques they are far less transparent to the user than the Census methods. Moreover, the *STM* method has the additional disadvantage that for each series a model has to be specified.

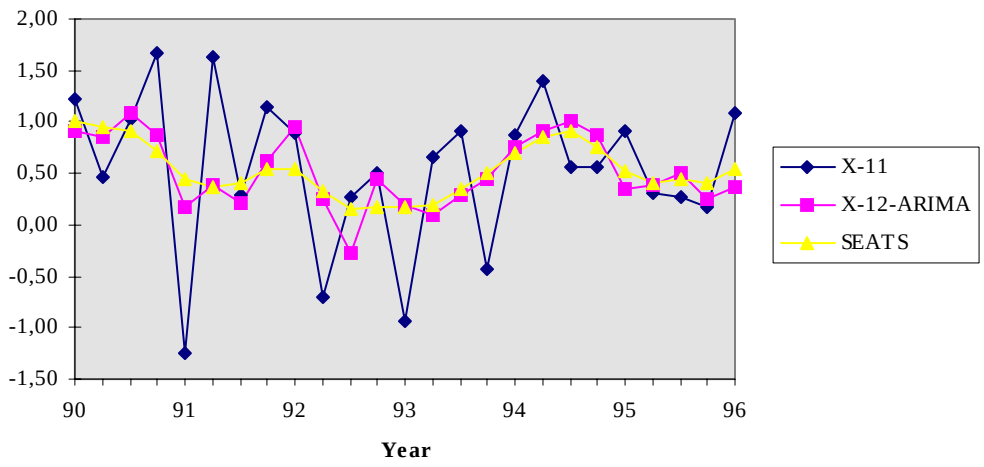
4. Performance of adjustment methods

Recent extensions of the Census methods can be seen as a compromise between the need for large scale applications of a transparent methodology that is simple and easy to use and the

theoretical advantages of the model-based methods. *X-12-ARIMA* also enhances the linearization of time series by removing causal effects, outliers and 'level-shifts' in the data before applying the fixed-filter methodology for the decomposition. Therefore, differences in the performance as between *X-12-ARIMA* and e.g. *SEATS* should be attributed merely to the different principles underlying the methods of decomposition. The Census methods pre-suppose that seasonal factors evolve slowly over time and that this can be adequately captured by taking moving averages. In essence *SEATS* also generates moving averages but these are derived from a different principle. The *SEATS* methodology directly aims at maximising the smoothness of the trend-cycle and seasonal component.

As an example the results of applying *X11*, *X-12-ARIMA* and *SEATS* to the series for GDP of the Netherlands are compared in figure 2. The figure gives quarterly growth rates computed from the seasonally adjusted series. In the *X11* results, causal effects - in particular calendar effects - have not been taken into account. The results from *X-12-ARIMA* and *SEATS* are derived from the same prior adjusted series and thereby reflect the differences between decomposition methods. Some differences are worth mentioning. Firstly, removing causal effects prior to seasonal adjustment gives a clearer picture of the business cycle: see the differences between *X11* and *X-12-ARIMA*. Secondly, in replacing the Census method of decomposition by the 'model-based' decomposition, an additional smoothing can be obtained. These results indicate the benefits of applying recently developed adjustment methods as compared to the old-fashioned *X-11* methodology: on the whole the new methods enhance a better interpretation of statistical results.

Figure 2: growth rates of GDP



References

Balestra P., 1987, Seasonal variation. In: J. Eatwell, M. Milgate, and P.·Newman (eds), The New Palgrave: Econometrics. Macmillan, London, pp. 198-200.

Balk, B. M., 1995, “Axiomatic Price Index Theory: A Survey”, International Statistical Review 63, pp 69-93.

Den Butter F.A.G., and M.M.G. Fase, 1991, Seasonal adjustment as a practical problem. North-Holland, Amsterdam.

Gomez, V. and Maravall, A., (1996), "Programs TRAMO and SEATS", Instructions for the User (Beta Version: September 1996), Working Paper N 9628, Bank of Spain

Harvey A.C., 1989, Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge.

Monsell B.C., 1988, Supplement to Census Technical Paper no. 15: The uses and features of X-11.2 and X-11Q.2. Bureau of the Census, Washington.

Shishkin J., A.H. Young, and J.C. Musgrave (1967), The X-11 variant of the Census method II seasonal adjustment program. Bureau of the Census, Technical Paper no. 15. Washington, D.C.

PART E PUBLICATION AND DISSEMINATION

The data set, resulting from the processing and analysis stage, comprises all aggregated information, available from the survey and subsequent processing and analysis operations. This set can be considered a *data warehouse*, containing the lowest possible aggregates which allow for publishing. These elementary aggregates are the building blocks for tabulations in all kinds of publications.

The actions to undertake in the final stage are no longer directed towards adding new information or improving reliability. Part of these actions, i.e. disclosure control measures, even result in the opposite, in that they suppress information.

The essence of the final stage is to structure, present and distribute the data produced in such a way that a maximum number of users is reached with information fitting their needs best. Before dealing with publication and dissemination strategies, we will give a brief outline of some common rules of game, setting limits with respect to the freedom of movement of NSI's on the information market.

I. THE TASK OF NATIONAL STATISTICAL INSTITUTES

Author: Ad Willeboordse

1. Introduction

In general the task of NSIs can be described as "the collection, processing and publishing of all statistical information which is considered useful for science, practice and government policy". From such a description one cannot immediately determine the boundaries within which an NSI is supposed to operate. Still, these boundaries should be clear-cut before developing a publication and dissemination strategy. In this introducing chapter we will provide a brief overview of boundaries commonly applied by western NSI's.

2. Delineation of NSI tasks

The boundaries of NSI tasks with respect to the nature of their product can be delineated by three benchmarks: confidentiality, equality and objectivity:

- The first and foremost constraint is that an NSI will never publish data that would lead to disclosure of information regarding individuals, institutions or businesses. In business statistics, a commonly accepted rule is that a tabulation cell should comprise at least 3 units, and that, for cells with larger numbers, the three units with the largest values should, together, do not dominate, i.e. account for less than 70 % of the cell value.

  It is, in principle, possible to remove this rule in individual cases by requesting the dominating respondent(s) to authorise the agency to disclose. In general, one should only apply such an escape in important cases. In the next chapter methods are described to control the risk of disclosure.

- Statistics compiled by NSI's are collective goods. This implies that every citizen can take note of statistical data under equal terms, and that no users are privileged. In this respect it is especially important to ensure that no new data are supplied to anyone before the moment of official publication, as well as to guarantee that results from a survey are available to anyone from the moment the first publication is issued. In most cases the press release is the first publication. This release not only serves as a data supplier, but also as a signal for the reader that additional information can be obtained by telephone or fax.

- Issued data should not be accompanied by judgements or recommendations; the independent and objective position of NSIs does not permit subjective interpretations.

Crossing these borderlines bears the risk of a disturbed relationship with both of the parties an NSI heavily depends on. *Respondents* would become suspicious with respect to the protection of their privacy and *users* would become suspicious regarding the independence of the Office and subsequently the objectiveness and thus reliability of the data.

II. PUBLICATION AND DISSEMINATION STRATEGY

Author: Ad Willeboordse

1. Introduction

On several occasions we have stressed the fact that the design and implementation of a particular survey may never be an isolated action of a particular surveying unit within the NSI organisation. In the choice and the definition of variables and populations one has to take into account the standards and norms of the general framework. Data collection should be organised within the context of a corporate and balanced surveying strategy, directed towards minimising the response burden.

In the same way, publication and dissemination of the results of a particular survey should fit to the notions of a corporate publication and dissemination strategy, aiming at maximum user satisfaction.

Complying with a total strategy implies in the first place the notion that “areas of interest” of users are not necessarily delineated according to the scopes covered by individual surveys. Secondly, there is a wide variety of user groups and consequently a wide variety of partly overlapping areas of interest. The first notion requires publications that contain data which are derived from different surveys. The importance of co-ordinated concepts, as derived from the coherent framework, is proven at this point in the statistical process particularly. The second notion asks for a wide variety of “publications”, each of them partly overlapping as to their information contents. It is the attainments of modern information technology, which enables the NSI to meet this variety of user needs by a pre-eminently flexible response.

This chapter provides a brief overview of dissemination and distribution modes, applying when trying to satisfy the adage: different users need access to different data clusters by means of different data carriers.

2. Data carriers

Apart from facsimile transmission and oral supply of information (mainly by phone), data carriers can be divided into printed publications and electronic dissemination. The first vary from regular publications in standard format to incidental print-outs on request and letters. Electronic dissemination makes use of diskettes, tapes and on-line transmission. The data carrier applied depends on several factors. Wishes and possibilities of users are of course most important. Only part of them dispose of electronic facilities, but this part is rapidly growing. Other factors are the speed at which the data become available and the suitability of electronic data carriers for the transmission of bulk data. Because of their future importance we will go in some more detail with respect to the possibilities of on line access of electronic data banks.

3. Electronic dissemination: tailor-made and topical

The amount of statistical information available is immense. As data communication networks (Interment) are getting more popular every day, the demand for instantly available statistical information, retrieved by on line access, is rising fast.

Three types of clients can be distinguished:

- the occasional client who wants some basic figures;
- the client interested in a specific set of information on a regular base;
- the client who needs large amounts of data with changing needs.

The *first client* can find answers to his questions in the free information pages provided by public information systems in which data suppliers of any kind can participate. He will use his computer and modem and make contact with the data supplier, search through hierarchical menus or 'click on buttons' with his mouse, and transfer the information desired back to his own system.

The *second client* can take out a subscription on the needed data. Each time a renewed set of data is available, it is sent to his mailbox automatically.

The *third client* wants to search through large amounts of data to find what he can use. Performance, completeness and flexibility in producing his own statistical tables are key issues for him. This client benefits from a powerful tabulation tool with datacom features to reach for the latest data available. The statistical institute is the natural supplier of such a tool, as it can be used internally for answering inquiries made by phone.

In the field of electronic media, the ultimate and literally global outlet for statistical information is the *Interment*. The use of this new and extremely powerful dissemination tool is presently in its infancy. Still, it is clear that an NSI cannot afford to ignore it. The big challenge is not so much the supply of data as such, but rather the organisation of the supply in such a way that users can find their way and data from various sources are consistent. In this respect, meta data are of crucial importance. They explain the concepts used, they elaborate the processes and methods applied for collecting and compiling the data, and they guide the user through the labyrinth of information.

4. Central information desk

However important paper and electronic dissemination may be, in practice most purchases of information start with a letter or telephone call by the potential user. The number of occasional telephone and fax enquiries is increasing constantly. The same goes for information requests referring to data from different survey sources. In order to prevent clients from being sent from one desk to another, and in order to supply consistent information, it is useful to concentrate first line information service in a central information desk. The staff of this desk answers requests rapidly with the help of the central data base mentioned earlier.

5. Tailor-made information

The creation of a central information desk and the increased attention to electronic dissemination considerably endorse the policy to supply tailor-made information. On-line systems are tailor-made by definition because consultation and selection of information is up to the user. In printed publications readers can be explicitly invited to contact the statistical agency if more detailed information is needed. The size of publications can then be reduced while at the same time service to the clients is improved. This also enhances efficiency, as the circulation of most printed publications (subscriptions plus sales) is low. This is especially true for publications, mainly consisting of detailed tabulations.

Interpreting the data correctly is a main problem using these media. Instant availability of the appropriate meta-data is a fundamental requirement.

6. Summary

The value of a survey does not depend on the amount and quality of data *produced*, but eventually on the *use* that is made of them. An adequate publication and dissemination policy applies in order to maximise use. Next to traditional paper publications, advanced electronic media supply substantial facilities, in particular with respect to topicality and tailor-made information. This enables users to specify information requests going beyond data resulting from one particular survey. Consistency of concepts and data derived from different surveys is therefore of utmost importance. A coherent framework of business statistics, as presented in PART A, is the obvious instrument to attain this goal.

III. DISCLOSURE CONTROL OF TABULAR DATA

Author: Cor Citteur

1. Introduction

Results of surveys among enterprises are usually published in the form of tables. Microdata sets with data from enterprises are hardly ever published. The reason for this is that enterprises can easily be identified on the basis of only a few characteristics as the number of employees (in size classes), the main economic activity of the enterprise and the region. Therefore, most statistical offices do not publish microdata sets with information from enterprises. This was one of the conclusions from a survey held among a number of statistical offices, see Citteur and Willenborg (1993). For instance in The Netherlands, a law prohibits the dissemination of business microdata collected under this law. Therefore, we will concentrate on the statistical disclosure control (SDC) of tabular data.

In general, tables do not contain information from individual respondents, but aggregated information referring to a number of respondents. However, situations may occur in which it is possible to deduce information corresponding with an individual respondent from the aggregated total, e.g. when the contribution of one respondent dominates this total. In such a case an SDC measure is necessary. In this chapter we will briefly discuss the question when an aggregated total is considered too sensitive to be published and we discuss three possible SDC methods to protect the sensitive information. These methods are:

- modification of the classification scheme,
- suppressing of the sensitive cells, and
- rounding of cell values.

2. Sensitive cells in tables

The first step in the statistical disclosure control of tabular data is the identification of the sensitive cells, i.e. the cells that tend to reveal too much information about an individual respondent. A common way to determine whether a cell is considered sensitive is by means of a *dominance rule*. This rule states that if the sum of the contributions of n or fewer respondents accounts for more than k % of the total cell value than this cell value cannot be published. The values n and k in this formulation are parameters whose values have to be chosen. For example, one could choose $n = 3$ and $k = 75$.

Example

Take n=3, k=75%. Suppose the largest three turnover values in a given total of 2.500 are 1 000, 500 en 475 respectively (all in millions of guilders). The sum of these three amounts to 1 975, that is 79% of the total of all turnovers. The dominance rule at issue is violated thereby.

The main idea behind this dominance rule is the following. If a cell value is dominated by the value of *one* respondent, then his contribution can be estimated fairly accurately. In particular, if there is only one respondent, his contribution will be disclosed exactly. If the value of a cell is dominated by the contribution of *two* respondents (for instance, if these respondents take up 90% of the total cell value), each of these respondents is able to estimate the value of the contribution of the other one. In particular, if there are exactly two respondents, each of them can disclose the contribution of the other exactly by subtracting their own contribution from the total cell value. In general, if there are m respondents then m-1 of them can, by pooling their information, disclose information about the value of the data of the remaining respondent. The value n should therefore be chosen larger than the maximum size of (imagined) coalitions of respondents. In view of the previous remarks, as a minimal value $n = 2$ is recommended. This value means that a cell total should be based on the contributions of at least three respondents.

Apart from the dominance rule other rules have been suggested to determine sensitive cells as e.g. the *prior-posterior rule*. With this rule it is assumed that every respondent can estimate the contribution of any other respondent to within q percent of its respective value. After a table has been published, the information of the respondents changes and they may be able to make a better estimate of the contribution of another respondent. A cell is considered sensitive if it is possible to estimate the contribution of an individual respondent to that cell to within p ($< q$) percent of the original value. We refer to Cox (1981) for a discussion of this rule.

Example (continued)

Take q=30% en p=20%. Look at the turnover values of example 1 and rank the individual values by decreasing magnitude. We have:

$$S_3 \equiv \sum_{i=3}^{N} x_i = 1000$$

So:

$$q \times S_3 > p \times x_1$$

According to the elaboration of the prior-posterior-rule, and in contrast with the previous dominance rule, this fact shows that it is allowed to release the total of turnovers in question.

3. Statistical disclosure control measures

It is possible that relatively many cells in a table appear to be sensitive. In that case it is recommended that first a global disclosure control measure is taken like the *aggregation* of (some of) the spanning variables of the table. In this way the detail of the statistical information is reduced. Our aim is that the number of sensitive cells decreases when less detailed information is released.

Example

Consider the following table on turnover by region and economic activity;

((s) means sensitive).

Activity	1	2	3	4	Total
Region:					
A	40 (s)	40 (s)	30	70	180
B	30 (s)	5 (s)	25 (s)	10	70
C	12 (s)	25 (s)	203 (s)	60	300
Total	82	70	258	140	550

There appear to be many sensitive cells in this table. We decide therefore to collapse the columns of activity codes 1 and 2. We assume that the newly formed totals are no longer sensitive.

Activity	1 + 2	3	4	Total
Region:				
A	80	30	70	180
B	35	25 (s)	10	70
C	37	203 (s)	60	300
Total	152	258	140	550

After the aggregation of the variables one has to check again whether the table contains any sensitive cells. In general, not all sensitive cells will be eliminated after the collapsing of some of the rows and/or columns in the table. Depending on the number of remaining sensitive cells further measures are necessary. When still relatively many sensitive cells remain in the table, then a further reduction of the detail of the information in the table by *collapsing* other rows and/or columns is recommended. When only a few sensitive cells remain, it is recommended to take a local measure like the suppression of the sensitive cells. With this method the sensitive cells are not published.

***Example (continued*)**

In the previous example we were left with two sensitive cells. We can use the method of suppression for these.

Activity	1 + 2	3	4	Total
Region:				
A	80	30	70	180
B	35	X	10	70
C	37	X	60	300
Total	152	258	140	550

The suppression of a cell because the contents of this cell is considered sensitive according to e.g. a dominance rule is called primary suppression. Primary suppression alone is in general not sufficient to obtain a table which is safe for release. In a table the marginal totals are often given as well as the values of the cells. A cell which has been suppressed can then be computed by means of these marginal totals. Therefore, other cells have to be suppressed in order to avoid this possibility. This is called secondary suppression. Secondary suppression can be done in many different ways. Two aspects play an important role:

i. the sensitive cells should be adequately protected by the choice of the secondary suppressions; the ranges in which the values of the suppressed cells lie, should not be too small.
ii. the loss of information due to the secondary suppressions should be minimal.

Example (continued)

The previous table is an illustration of this phenomenon. Working within the rows of B and C it is a trivial task to recalculate the suppressed values. In order to overcome this problem we put some additional x in other cells. One choice for these is presented in the table below.

Activity	1 + 2	3	4	Total
Region:				
A	80	30	70	180
B	35	X	X	70
C	37	X	X	300
Total	152	258	140	550

Although it may be impossible to compute the exact values of the suppressed cells after secondary suppression, it will always be possible to compute the ranges in which the values of the cells lie, when it is for instance known that the values of the cells are all non-negative (cf.

Geurts, 1992). If these ranges of feasible values are small, then an attacker is able to obtain good estimates for the values in the suppressed cells. Therefore, secondary suppression must be done in such a way that the ranges in which the values of the suppressed cells lie are not too small. In other words, the secondary suppressions should provide sufficient ambiguity about the values of the sensitive cells in the table.

Example (continued)

Activity	1 + 2	3	4	Total
Region:				
A	80	30	70	180
B	35	0-35	0-35	70
C	37	193-228	35-70	300
Total	152	258	140	550

Different choices of the secondary suppressions will lead to a different loss of information. Usually the secondary suppressions are chosen such that a certain criterion function is minimised, where the minimization is performed over all possible choices of the secondary suppressions that adequately protect the sensitive cell values (see i)). Examples of the criteria to be minimised are the total number of suppressed cells, the number of respondents whose data are suppressed or the total value of the data suppressed. Also combinations or other criteria are possible. The choice of the criterion function which should be used, will depend on the purpose of the table.

To find secondary suppressions that satisfy both the aspects i) and ii) is in general a difficult matter. From an SDC point of view it is necessary that at least the first aspect is satisfied. Heuristics have been developed which yield secondary suppressions that satisfy i) and that approximately satisfy ii). We refer to Geurts (1992) and Kelly (1990) for a further description of the secondary cell suppression problem.

A third disclosure control measure to safeguard tabular data is *rounding* the cell values to integer multiples of a fixed base. Rounding can be seen as a special way of perturbing the data. In both cases the data are modified in order to protect the underlying information.

Several methods exist for the rounding of tabular data. The easiest one is conventional rounding where each value is rounded to the nearest multiple of the rounding base. Conventional rounding has two serious limitations. First, the additive structure of the table is usually not preserved, i.e. the sum of the rounded cell values is not necessarily equal to the rounded sum of the cell values. A second and from an SDC point of view important constraint is that it is sometimes still possible to disclose the original cell values after rounding.

In order to overcome these limitations so-called controlled (random) rounding procedures are suggested in Fellegi (1975) respectively Cox (1987) for one- respectively two-dimensional tables. With both procedures it is possible to preserve the additive structure of the original table. This is achieved by rounding the cell values to one of the two adjacent multiples of the rounding base instead of rounding the values to the nearest one as is done with conventional rounding. To which of the adjacent multiples a value is rounded is determined by a random process. Due to this process the procedure is unbiased. The expected outcome of the rounded value is equal to the original value.

Example

Rounding a one-dimensional table according to Fellegi. The original table stands to the left; two possible rounded versions are given next. The rounding is to the base of 5.

	Original	Rounded a	Rounded b
	12	15	10
	5	5	5
	3	0	5
	18	20	20
	23	25	20
T	61	65	60

Example

Rounding a two-dimensional table according to Cox. As in the previous example the rounding is to the base of 5.

Original				Rounded			
			Total				Total
9	14	4	27	10	15	5	30
8	10	6	24	10	10	5	25
19	12	7	38	15	10	10	35
T 36	36	17	89	35	35	20	90

A practical aspect with rounding is the choice of the rounding base. Rounding should provide the sensitive cells with a sufficient range of ambiguity. Unacceptable close estimates of those cells should not be possible. Therefore, the rounding base should be chosen larger than a certain minimum value which depends on the values of the sensitive cells. On the other hand, the rounding base should not be chosen too large, since in that case too much information regarding the nonsensitive cells may be lost.

No general techniques exist for the controlled rounding of higher dimensional tables. Cox (1987) gives an example of a simple three-dimensional table in which it was impossible to round each value to one of the two adjacent multiples of the rounding base and at the same time to preserve the additive structure of the table. Some heuristics to deal with three-dimensional tables have been developed at the U.S. Bureau of the Census (see Fagan et al., 1988). Also in Kelly (1990) heuristics are developed. For higher dimensional tables satisfactory heuristics seem hard to find.

4. More than one table

The previous sections deal with the disclosure control of one single table only. When a set of linked tables, i.e. tables with common variables stemming from the same microdata, are published, additional problems may arise. It is possible that a table in itself does not contain any sensitive cells, but that by combining the information of that table with information from the other tables it is still possible to disclose information about an individual respondent. In De Vries (1993) a method is described to determine which additional information can be derived from a set of linked tables. If unacceptable accurate estimates of the contribution of a respondent are possible, then an SDC measure apply. For instance, one could delete one or more of the tables from the set of linked tables. Another option is to protect the original microdata file against disclosure. In that way all tables which are based on these microdata set are immediately safe. This method may especially be useful when a large number of tables stemming from the same microdata set is published. A drawback of this method may be that a lot of information gets lost. Further research is still necessary on these points.

Example

Consider three 2 x 2 - tables, which pertain to the population of shopkeepers in a given city. Criteria for break down are sex, location and financial position. The latter is a sensitive variable; the score on it is not publicly known on a large scale. The opposite holds for sex and place of business.

A closer look at the interrelationships between these tables reveals that all male shopkeepers in the outskirts have a weak financial position, and that the financial position of all female shopkeepers in the centre is strong.

1. Shopkeepers by sex and location

Sex	location		total
	center	outskirts	
male	19	6	25
female	3	6	9

2. Shopkeepers by sex and financial position

Sex	financial position		total
	weak	strong	
male	13	12	25
female	3	6	9

3. Location and financial position of shopkeepers

Location	financial position		total
	weak	strong	
center	7	15	22
outskirts	9	3	12

5. Additional meta data

Another problem is that additional information about the SDC method used may lead to disclosure of sensitive information. For example, suppose that the following dominance rule is used: a cell is sensitive if at least 80% of the value is the combined result of the data of two companies. Suppose, furthermore, that all the companies are in the sample. Now, suppose that there are three companies contributing to a certain cell and that the data of the largest company constitutes 50% of the value of this cell. If the cell is published and the largest company happens to know the parameters of the dominance rule, then this company can deduce that the data of the second largest company is between 25% and 30% of the total value of the cell. On the other hand, if data are suppressed and the largest company again happens to know the parameters values, then the largest company can deduce that the data of the second largest company constitutes between 30% and 50% of the total value of the data. Absolute secrecy about the parameters of the dominance rule is the first step to avoid these problems.

References

Citteur, C.A.W. and L.C.R.J. Willenborg, 1993, Public use microdata files: current practices at national statistical bureaus. Journal of Official Statistics, Vol. 9, pp. 783-794.

Cox, L.H., 1981, Linear sensitivity measures in statistical disclosurecontrol, Journal of Statistical Planning and Inference, Vol. 5, pp. 153-164.

Cox, L.H., 1987, A constructive procedure for unbiased controlled rounding, Journal of the American Statistical Association, Vol. 82, pp. 520-524.

De Vries, R.E., 1993, Disclosure control of tabular data using subtables, Report, Statistics Netherlands, Voorburg.

De Waal, A.G. and L.C.R.J. Willenborg, Principles of statistical disclosure control, Report, Statistics Netherlands, Voorburg, (submitted for publication).

Fagan, J., B. Greenberg and B. Hemming, 1988, Controlled rounding of three-dimensional tables, Statistical Research Division Report Series, Bureau of the Census, Washington D.C.

Fellegi, I.P., 1975, Controlled random rounding, Survey Methodology, Vol. 1, pp. 123-133.

Geurts, J., 1992, Heuristics for cell suppression in tables, Report, Statistics Netherlands, Voorburg.

Kelly, J.P., 1990, Confidentiality protection in two- and three-dimensional tables, Ph.D. thesis, University of Maryland, College Park, Maryland.

Kish, L. Survey Sampling, John Wiley, New York (1967)

Särndal, C.E., Swensson, B., and Wretman, J., Model Assisted Survey Sampling, Springer-Verlag (1992).

IV. PROFILE OF A STANDARD PUBLICATION

Author: Ad Willeboordse

1. Introduction

Design, structure and contents of publications heavily depend on the subject and the user groups. Qualified expert users are more interested in detailed tables, guided by relevant explanatory notes, and less in illuminating text and graphs, whereas for the general public easy readability and accessibility are important. Still, some guidelines apply, in particular with respect to information on the survey method as well as tabulation design.

2. Meta data

Users have the right to be informed on the characteristics of the product they receive. This information concerns three aspects:

1. information on the meaning of the statistical *concepts* used in the tables. This usually takes the form of definitions. Explanation should not be confined to the variables but relates to the statistical units, the classifications used and the populations as well

Example

The publication on turnover in the computer services industry should explicitly make clear:

- what exactly is meant by *turnover*. A definition listing the components of the concept (inclusions and exclusions) is often more informative than a more theoretical definition. The ANNEX to the METHODOLOGICAL MANUAL OF STRUCTURAL BUSINESS STATISTICS provides useful definitions of a large number of concepts;
- which *unit type* is used and how it is defined. Mentioning of the unit type gives the user some indication of secondary activities being included or not, and whether ancillary units are involved or not;
- which *classification rules* have been applied;
- how the *population* is delimited. E.g., If the informal economy is excluded from the data, this should be explicitly mentioned.

2. information on the *collection* and *processing* of the data “produced”;

Example (continued)

- which *collection method* has been used (paper, telephone etc.)?;
- what was the *scope* of the collection?. If the data were based on a cut off (e.g. more than 20 persons employed) and the data for the smaller units were calculated, estimated or imputed, this should be explicitly mentioned;
- how was dealt with non-response?
- how have the data been edited?
- which methods have been used for weighting, seasonal adjustment, deflation, time series etc.?

Information on collection and processing is in particular interesting for users because they implicitly give some indication of the quality. Besides, there are also quantitative quality measures.

3. data on the *quality* of the data.
   As these are often included in the tables, they will be dealt with in the next section.

3. Tabulations

Introduction

Statistical tables are the heart of a publication. As we have seen in PART A, these tabulations have to be designed already in the beginning of the survey design process, where they constitute the *table outline.*

A first and foremost condition for each table is that the message to communicate can be easily understood. The intended data should be presented clearly and unambiguous and the title should describe essentially the contents of the table. The wording must be as informative as possible and easy to read and understand. This also holds for the stubs explaining the components of the table. If necessary, shortened versions of title or stubs can be used in combination with footnotes.

Surveyed data

A table commonly consists of cells arranged in rows and columns. In case of a sample survey the cell contents are usually estimates of totals or percentages of a predefined population. The number of digits in which the cell values are expressed should not give a false impression of the statistical reliability of the data. The policy of the statistical agency concerning confidentiality may restrict the number of digits per cell or it may be even a reason to suppress certain cells.

Rounding is often carried out to remove irrelevant digits. This is done for readability, reliability or for confidentiality reasons. The title or the stubs should indicate that the cell values are rounded, for example, to the nearest thousand in case of estimated totals or to one decimal place in case of estimated percentages.

Tables produced by statistical offices are often copied from their original publication and inserted into other texts. Therefore they should be self-explaining and as much as possible designed to stand on their own, leaving no room for misinterpretation.

Quality indicators

An important issue for sample surveys are precision indicators. It must be decided whether estimates of variances, standard deviations, confidence intervals or coefficients of variation will

be published. The presentation can be done in different ways, in the table itself or in a separate part of the publication. The last option has the disadvantage that the reader has no overall picture of estimates and their precision within one table. On the other hand, providing all estimates together with a measure of their precision can complicate the presentation of the tables. Statistical agencies usually try to hit upon a compromise. Statistics Canada often uses a simple alphabetic notation (A, B, C, ...) to express the quality of individual table entries.

The presentation of the quality of the intended output should be part of the table outline for a sample survey. A chapter on this topic can be found in Kish (1967). More general remarks on quality declarations in statistical publications, with some other references to technical literature, are given in Särndal, Swensson and Wretman (1992).

Automation aspects

Dedicated tools like Abacus, Manipula, SAS, SPSS and Stata are all able to aggregate and tabulate statistical data, although they considerably differ as to the available functions, and the ease with which it is possible to exchange data and meta data with other tools. When it comes to choosing one of these tools, availability, price and personal flavour come into play. Physical constraints as maximum size of the data set and processing speed are important as well. A last but certainly not negligible determinant is the required lay-out quality of the statistical tables.

4. Timeliness

The speed of dissemination is negatively affected by the accuracy and detail required and the time it takes to produce the publication, as described above.

Supplying the principal data by means of press release and weekly statistical bulletin gives some relief. Apart from pursuing the earliest possible release of information, one may publish *provisional* data, which are subject to revision. Actually, a market-oriented policy requires that release of information before or on a certain date be guaranteed. This can be achieved by pre-announced release dates. Moreover, users appreciate predictability of release of information.

5. Other features

Tables are often accompanied by text and graphs for reasons of accessibility. Although accessibility partly depends on the nature and detail of the information itself, there are ways to enhance it. Clear or even popular language, transparent lay-out and illustrations such as colour-graphs all increase the readability of publications. However, accessibility is not the ultimate goal of statistics: some users may need explicit attention to the ins and outs of the data. A straightforward aspect of accessibility is language. Due to increasing internationalisation, for non-Anglo-Saxon countries it is recommended to issue at least the outline of contents and the major data in English.

6. Closing the cycle

It is true that usage of advanced and multiple dissemination systems dramatically increases accessibility of data. However, eventually the degree of satisfaction of the user will mainly depend on the extent to which the information supplied fits his needs. In this respect comparison of performance with user needs as specified in the very first stage of the process is the first measure applying.

7. Reopening the cycle

However, user needs are dynamic. In particular after issuing the outcomes of a new survey, users will be inclined to reconsider their needs and priorities. Therefore, measuring user satisfaction as well as market research should be recurrent operations. It must be stressed that such monitoring is not only worthwhile to keep ahead with *expanding* user needs, but also to reconsider the usefulness of existing data surveyed. Balancing effort and returns is indeed a going concern. The fact that a certain data item is surveyed because it *has been* surveyed for years is a poor justification towards respondents that become ever more critical.

CONCLUDING REMARKS

Survey design and implementation is a chain of actions and interactions with three players involved: users, respondents and, in between, surveying statisticians. A successful survey is characterised by both a maximum compliance with the needs of the users, and a minimum burden on the respondent. It is the task and the challenge of the statistician to simultaneously satisfy these two, often conflicting, interests, and to produce statistical data that are comparable and consistent with those of other surveys.

These demands affect all stages of the survey design:

- While setting the survey objectives (PART A) user needs are "translated" in such a way that they fit as much as possible to the standards and concepts of the general framework, and that the gaps with reporting possibilities of respondents are not unbridgeable.
- Subsequently the survey objectives are operationalized in such a way (PART B) that the sample is as small as possible and the contents of the questionnaire are fitted to the possibilities of the respondent.
- Data collection (PART C) is organised in such a way that response burden is spread equally over respondents, questionnaires for different surveys are standardised or even integrated, and redundancy of questions is banned. EDI is an important tool to achieve these goals.
- During data processing (PART D) data from different survey sources are matched so as to obtain consistency, while processing methods (editing, imputation, estimation) are designed as uniform as possible.
- Publication and dissemination (PART E) is organised in such a way that the data resulting from the survey can be related easily with relevant data from other surveys. An on line data bank, which denies borderlines between surveys, and where information from different surveys is gathered in one table, is a powerful tool in this respect.

ANNEX ACRONYMS

CN	Combined Nomenclature
CPA	Classification of Products by Activity
CPC	Central Product Classification
CR SBS	Community Regulation on Structural Business Statistics
CR SBR	Community Regulation on Statistical Business Registers
CR STI	Community Regulation on Short term Indicators
CR LCS	Community Regulation on Labor Cost Structure
EDI	Electronic Data Interchange
ESA	European System of Accounts
EU	European Union
EUROSTAT	Statistical Office of the European Union
ISIC	International Standard Industrial Classification
KAU	Kind of Activity Unit
NACE Rev. 1	Statistical Classification of Economic Activities in the European Community
NSI	National Statistical Institute
PRODCOM	List of products to be used in production statistics
SBR	Statistical Business Register
SNA	System of national Accounts
TQM	Total Quality Management
TSD	Total Survey Design
UHP	Unit of Homogeneous Production
UN	United Nations

[i] For an illustration of the links between the three unit types, see PART A, Chapter .....

[ii] For an extensive treatment of a variety of strategies we refer to Cochran (1977), Hedayat and Sinha (1991), and Särndal, Swensson and Wretman (1992).