Statistics Netherlands

Research and Development Division
Department of Statistical Methods

P.O.Box 4481
6401 CZ Heerlen
The Netherlands

A model for evaluating the quality of business surveys

Robbert H. Renssen, Astrea Camstra, Christine Huegen,
Wim J.G. Hacking, René D. Huigen

Remarks:
This report is the final result of the project 'Data Quality' financed by Eurostat and carried out by Statistics Netherlands.

Project number: *RSM-30060*
BPA number: *6279-98-RSM*
Datum: *14 September 1998*

A MODEL FOR EVALUATING THE QUALITY OF BUSINESS SURVEYS

Summary

In this report we will describe a model that is developed to evaluate the quality of structural business statistics of the EU Member States in relation to the survey costs. Starting point is the survey process that determines the quality and costs of the statistics produced. By means of an activity-quality matrix and activity-costs matrices we link quality aspects, production costs and response burden to the activities in the survey process. In order to assess the quality of the survey results we derive two quantitative measures, one for the sampling error and one for the non sampling error due to non response. We also give a number of qualitative measures with respect to the non sampling errors. We define cost functions for each activity of the survey process taking into account that information on costs is difficult to obtain. The optimal survey strategy is calculated by minimizing the variance subject to the budget restriction. In evaluating the trade off between precision and costs, the quality and cost indicators are defined as a function of a limited set of instruments which can be used to simulate changes in the process characteristics. The model and the simulations are implemented in a computer program. A questionnaire was designed to collect the information needed from the Member States.

Keywords: quality, survey costs, response burden, optimum allocation

Table of contents

Introduction

Recently EUROSTAT has initiated the project ‘Data Quality’. The objective of the project is to find means to evaluate the quality of business statistics in the EU taking into account the costs of the survey process including response burden. The project must result in a software tool which can be used to decide what is the best and most cost efficient practice to obtain estimates with a pre-specified accuracy given the specific properties of the business surveys in the Member States. The main reason for starting this project is that the new Council Regulations concerning structural business statistics requires the Member States to inform the Commission about the quality of the statistics compiled. It is explicitly mentioned that sufficient information has to be provided about the accuracy of the survey results.

The project has been carried out by Statistics Netherlands in line with the project proposal (Statistics Netherlands, 1995). In order to compare and improve the quality of structural business statistics in the EU a model is developed relating quality to the costs of the business survey process. The model is expected to give us more insight in the differences in process characteristics between the Member States and how these differences affect the survey results. For the model to be of practical use, reliable information about the survey process of the Member States is required. When this information is available the model may serve as a useful starting point for discussion with the Member States on the quality of business statistics. This discussion has already started in the working group “Assessment of the quality of business statistics” which took place the same period this project was carried out. We have made thankful use of the results of the working group.

In Part A of this report the formulation of the problem is worked out by presenting a conceptual framework for modeling quality and costs of the business survey process in general. Following closely the EUROSTAT documents for the working group “Assessment of quality of business statistics” several quality aspects of statistical products are discussed in Section 1. In Section 2 we present a general framework to evaluate the quality of the survey results. This evaluation is done by modelling quality and cost indicators. Both types of indicators are based on the data, process, and costs characteristics that are obtained from the survey process. The statistical survey process is briefly discussed in Section 3. Based on the “Handbook on the Design and Implementation of Business Surveys” (EUROSTAT, 1998) we show that the survey process is a sequence of activities which together determine the quality of the statistical product. In Section 4 the quality aspects are linked to the activities in the survey process. Following the project proposal (Statistics Netherlands, 1995) we take accuracy as the quality aspect to be evaluated. In Section 5 we show for this quality aspect what costs including response burden are involved.

In Part B the model is described in detail. The data and process characteristics of a business survey form the input of the model, while the quality indicators referring to the accuracy of the survey results and the cost indicators constitute the output. The quality indicators are derived from the input by using familiar quantitative measures on the one hand and simple qualitative measures and aggregation rules on the other hand. Furthermore, the model relates costs to precision by simulating changes in the sample size using a limited set of instruments. In Section 7 we describe the transformation of the process characteristics into qualitative quality indicators for the non sampling errors. In Section 8 we formulate a simple model for response behaviour needed to derive several quantitative quality indicators with respect to the sampling error. In Section 9 we present a set of cost functions taking into account the information needed for the simulation. Given total costs as budget restriction the optimum allocation is calculated in Section 10. In Section 11 we focus on simulating changes in the gross sample size and the follow up intensity to illustrate the relation between costs and precision. In Section 12 we discuss the interrelation between sampling error and frame error as a starting point for a future extension of the model incorporating non sampling errors. In Section 13 attention is given to response burden as a cost indicator. In section 14 utility functions are introduced to interrelate the quality and cost indicators.

The model is implemented in a software tool. The program is documented in Part C. To illustrate the program an example / sample case is worked out. The case shows what kind of information is needed as input, how this information must be entered into the program, and what kind of output is generated. Besides this hypothetical case, some real cases relating to business surveys of the Member States are supplied with the software.

The report ends with a brief summary of the project results. Appendix A gives the proof of a certain aspect of the theoretical model used. The information that is required from the Member States about their survey process is listed in Appendices B, C and D. In Appendix E criteria are formulated that can serve as a guideline for the rather subjective translation of the answer categories into qualitative quality scores with respect to non sampling errors.

Part A. The conceptual framework

1. Quality

The regulation on structural business statistics requires the Member States of the European Union to produce a report on the quality of the statistics they provide. What do we mean by quality and how do we measure quality? The working group “Assessment of the quality of business statistics” has worked out in detail the concept of quality for statistical products. To measure the quality of statistical products several quality aspects can be distinguished. Following EUROSTAT (1997, 1998) we distinguish six quality aspects:

- Relevance,
- Accuracy,
- Timeliness and punctuality,
- Accessibility and clarity,
- Comparability,
- Coherence.

Relevance is the extent to which survey results correspond to the needs of the users. To be relevant statistical information must describe matters that are of interest for the user in terms that are recognizable and interpretable. Relevance depends strongly on a careful transformation of user needs in product specifications. The larger the discrepancy between realized output and user needs, the less relevant a statistical product will be.

Accuracy is a measure of the total error associated with a statistical outcome. The total error can be subdivided in the sampling error and non sampling error. We prefer to consider these two types of error as different quality aspects. Errors in the estimates only due to the fact that the estimates are obtained from a sample rather than a complete enumeration are referred to as sampling error. Because estimates are obtained from the net sample, i.e. observed sample, rather than the gross sample, i.e. designed sample, the reduction of the sample size due to non response is included in the sampling error. All other errors in the estimates arising at any stage in the survey process are referred to collectively as non sampling error. Important sources of non sampling errors are selective non response, improper coverage and measurement errors (Lessler and Kalsbeek, 1992).

Timeliness refers to the period between the reference period the statistical outcomes refer to and the moment the survey results come available to the users. To be of interest for users results must be available as soon as possible. Users want up-to-date outcomes which are disseminated on time at pre-established dates (punctuality). In practice there is a clear trade off between timeliness and accuracy, especially when provisional estimates are provided.

Accessibility and clarity of the information is a quality aspect dealing with the dissemination of statistical data. Users must have easy access to the data and know what kind of statistics are available. Moreover, an adequate documentation can help users to use and interpret the data in a correct way. This quality aspect will not be discussed further in this report.

Comparability is a quality aspect of interest when statistics are compiled for EUROSTAT. To get comparable results at the Community level the business surveys of the Member States should be coordinated with respect to the definition of the population, the type of statistical unit, the definitions of the variables, the level of accuracy, and so on. In our opinion comparability at Community level is determined strongly by the quality of the survey results at National level. In this context it would be misleading to consider comparability as an additional quality aspect of the statistical product. Note that this view is fully in line with the intention of the project that means have to be found for controlling the quality of the statistics produced by the Member States in order to get comparable and reliable outcomes at the Community level.

As last quality aspect we distinguish coherence. Coherence refers to both the consistency at the conceptual level and the consistency of the statistical outcomes themselves. To avoid too much discrepancies in concepts the definitions and classifications used in different surveys have to be highly coordinated. Furthermore, inconsistencies in the outcomes from different surveys describing the same phenomena should be minimized.

Cost is a measure of performance reflecting the efficiency of the process. Following EUROSTAT (1997) we see cost not as a quality aspect, but as an additional factor to be associated with quality. Besides costs made by the statistical agency there are also costs involved for the enterprises by filling in questionnaires, the so-called response burden. Quality and cost together determine the total performance of the product. From this point of view it is the trade off between cost and quality that is of special interest for a statistical agency.

To evaluate the statistical product with respect to these quality aspects appropriate quality measures have to be defined. In practice the translation of quality aspects into concrete quality measures is rather problematic. In fact, quantitative measures of quality are relatively scarce and limited in use. Quality measures like satisfaction scores, variance estimates, non response rates, delivery time, and number of detected inconsistencies do only partly cover the quality aspects in question. The problems with quality measures are most severe for the quality aspect accuracy, especially when dealing with non sampling errors.

Normally the size of (sampling) errors cannot be quantified because the true value to be estimated is unknown. However if an unbiased or an approximately unbiased estimator is used, which is often the case, a convenient measure of the sampling error is the variance (or standard error) of the estimator. This measure of precision of an estimator can be estimated from the sample. Unfortunately, variance due to non sampling errors, like measurement errors, is also included in the estimate of the sampling variance. So, in practice it is difficult to distinguish sampling errors from non sampling errors. The diversity and complexity of non sampling error sources makes it even more difficult to evaluate total error as a measure for accuracy. Sources of non sampling error are often evaluated separately. Typically these evaluations provide no more than an indication of the potential for bias and variance resulting from an error source.

In practice we often see that quality assessment is done by measuring the quality of the statistical product and process in more qualitative terms. This is also suggested by the working group “Assessment of the quality of business statistics” in their proposal for a standard quality report (see EUROSTAT, 1996a). A careful inspection of the survey process, i.e. the methodology applied and the data sources used, often gives useful information about the quality of the survey results. Giving every activity a quality score one gets a clear picture of the strong and weak parts of the total process. Of course, assessment in qualitative terms is less objective and always open to question. But if quantification of quality gives too many problems, the use of qualitative measures is the only alternative.

2. General framework for quality evaluation

In order to get comparable and reliable outcomes for business surveys from different statistical agencies, means have to be found for assessing the quality of the survey results in relation to the survey costs. As we have seen in Section 1, quality assessment is not easy to perform in practice. The quality of statistical data can not be determined from the data itself. Indicators for the quality of survey results, including well known measures as variance and response rates, can only be obtained by monitoring the underlying survey process and data sources in detail. For evaluating the costs it is also needed to go back to the process. The main object to be studied is therefore the statistical survey process. Having information on the process and the data sources used is vital for any form of quality evaluation.

Figure 1 gives the general framework we use to assess quality. The survey process results in statistical data, which are often presented in tables. The quality of the survey results strongly depends on the performance of the process, i.e. the effectiveness of the methodology applied, as can be determined from the data and process characteristics. The impact of the survey process on the quality of its results and on the total costs are described by means of a model. The input of this model with respect to the survey process consists of a number of data and process characteristics, including costs. The output of the model are quality and cost indicators. In the model all quality indicators will refer to accuracy as the quality aspect to be evaluated. Note that we see statistical integration, i.e. the process where the survey results are aligned with other survey results, as an additional mean for quality control.

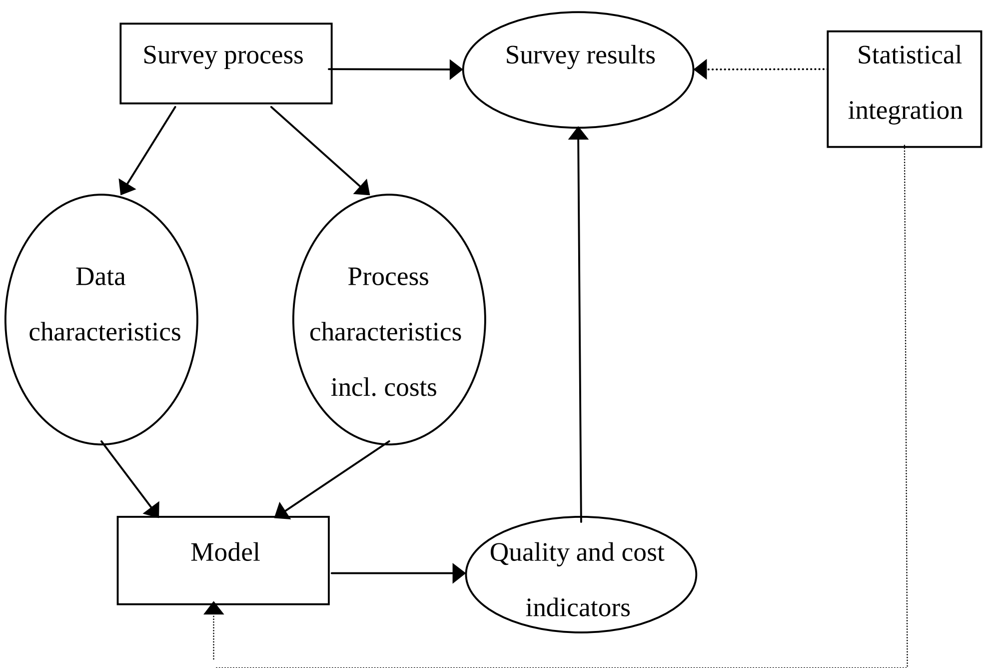


Fig. 1. Scheme of the general framework

Data characteristics are needed to calculate the standard error, which is modeled as a quantitative indicator for the sampling error. We distinguish the following data characteristics (see also Appendix C):

- Stratum net sample size,
- Stratum sample totals/sample means
- Stratum sample variances,
- Stratum sample multiple correlation coefficients.

For reasons of simplicity we assume that all business samples will be drawn according to a stratified design. Furthermore, we assume that each stratum total is estimated by a generalized regression estimator based on multiple regression models (see Särndal et al., 1992). The way the data characteristics are used in the model will be discussed further in Section 8.

In Section 3 we will see that the statistical survey process is a sequence of activities. In principle each of these activities can serve as instruments to alter the quality or total costs of the statistical results. For example, there are four possibilities to realize data collection, namely by mail, telephone, interviewer or electronic. Those four options formulate the instrument "data collection method" and a specific choice "mail" characterizes this activity of the process. Every activity/instrument is associated directly with one or more process characteristics influencing quality and/or costs. An inexhaustive list of process characteristics linked to the various survey activities is given in Table 1. In Section 7 a more complete list of process characteristics is given.

Table 1. Process characteristics grouped by activity/instrument (example)

Activity/Instrument	Process characteristics
Use and maintenance of frame	• Size and frequency of quality survey • Number of dead or misclassified units
Sampling design	• Gross sample size • Number of strata • Neyman/proportional/other allocation
Questionnaire design	• Efforts made for quality testing • Mean time to fill in questionnaire
Data collection	• Mail/telephone/interviewer/electronic
Follow up procedures	• Number of recalls per stratum
Data entry	• Input by hand/automatically
Editing method	• By hand/automatically/both • Micro/macro/selective editing
Estimation procedure	• Use of auxiliary variables

We distinguish between two types of quality indicators, namely quantitative and qualitative indicators. The quantitative indicators, like variance and response fractions, relate particularly to the sampling error. The qualitative indicators refer to the non sampling errors and will be presented as rough scores on the process instruments.

As stated in the project proposal (Statistics Netherlands, 1995) the intention is to build a model by which quality effects of alternative survey practices, i.e. alternative process characteristics, can be simulated. A first form of quality evaluation is to compare quality indicators of similar surveys with each other. As every Member State will use its own methodology and data sources, one might expect differences in quality due to differences in the data and process characteristics. A careful analysis of these differences can give insight in the possibilities for the Member States to adjust their survey process to improve quality. The qualitative quality indicators we will present in section 7 are especially useful for this form of comparative analysis.

A second form of quality evaluation is to quantify the effects on quality and costs of alternative survey practices for one specific survey process. By choosing only one activity as an instrument and keeping all other activities unchanged, the impact on accuracy and costs of the instrument under consideration can be examined taking into account the specific properties of the survey in question. In this way the trade off between accuracy (precision) and costs can be quantified giving the Member States a clear picture on the possibilities and costs of quality improvement. In Sections 10 and 11 we simulate the effects of several alternative survey practices. The following process characteristics will serve as an instrument: gross sample size, number of recalls, allocation scheme, and the use of auxiliary information respectively.

Before we specify the model further, we will first discuss the relation between survey activities, quality aspects and costs. An accurate description of the impact the survey process has on quality and costs is fundamental for our approach of quality evaluation.

3. The statistical survey process

Generally speaking, the objective of statistical surveys is to collect information about a finite population. The set of activities needed to design and implement a statistical survey is called the statistical survey process. Following EUROSTAT (1998) we distinguish five phases:

- Setting the survey objectives,
- Preparing the survey operations,
- Sampling and data collection,
- Data processing and analysis,
- Publication.

The very first activity when considering a survey is the formulation of the need for statistical information. The needs have to be translated into a detailed specification of the intended statistical output. This implies a careful description of the variables of interest and the corresponding population parameters, including the choice of statistical units and classifications. After formulating the statistical needs one should explore secondary sources by which these needs (partially) may be fulfilled. In this context, we note that there have been new developments with respect to electronic information technology, one of the most promising of them being Electronic Data Interchange. The use of EDI will increase the amount of secondary sources considerably. If the statistical needs are covered insufficiently by these sources one may consider to perform a sample survey.

Generally speaking, the objective of a sample survey is to collect information by means of a sample. There are two lines of preparatory operations, namely the design of the sample (including the choice of the sampling frame) and the design of the questionnaire. The population to be sampled should coincide with the population about which information is wanted. Sampling theory provides a wide variety of methods according to which samples can be drawn. The choice of a specific design depends on the registered units and auxiliary information in the sampling frame. No rules of questionnaire design are generally valid. However, in EUROSTAT (1998) some guidelines are mentioned.

In the third phase the sample is drawn from the sampling frame according to the sampling design. The sampled units are informed by sending an advance letter. Next, they are approached either by mail, telephone (including CATI), interviewer (including CAPI), electronic way (tape, diskette, EDI), or some mix between them. Follow up procedures are performed in order to enhance response and data quality. The collection process ends with entering the data reported in a data file, comprising the records of all (responding) sampled units.

The activities which can be distinguished in data processing are editing, imputation, estimation (raising, re-weighting), and statistical integration. Shortly, editing and imputation are activities to clean up raw data and weighting is an activity to obtain certain estimates from the cleaned data. Often, external sources are utilized to perform these activities. Statistical integration is an activity to confront the obtained estimates with corresponding estimates from different surveys, and if necessary to adjust them. Besides weighting procedures, the (cleaned) data can also be analyzed by applying statistical techniques, such as correspondence analysis, regression analysis, index numbers, time series, and so on.

The final phase comprises the making of publications, of which the tabulations and analysis results are the main features, and disclosure control techniques play an important role. The aim of statistical disclosure is to limit the risk that sensitive information of individual respondents can be disclosed from a data set. The activities which can be distinguished in a statistical survey process are summarized below.

Setting the survey objectives

- Formulating user needs
- Translation output specification
- Use of secondary sources

Preparing the survey operations

- Sampling design, including choice of sampling frame
- Questionnaire design

Sampling and data collection

- Drawing the sample
- Data collection
- Follow up procedures
- Data entry

Data processing and analysis

- Editing and imputation
- Estimation procedures, use of external sources
- Statistical integration
- Analysis

Publication

- Tabulation and analysis results
- Disclosure control

4. The activity-quality matrix

In Section 3 a number of activities to make a statistical product are given. Each activity will affect one or more of the quality aspects discussed in Section 1. In Table 2 these quality aspects are related to the activities of the statistical production process. A specific activity only influences the quality aspects which are marked by means of a cross. For example, the choice of the data collection method will influence the sampling error, the non sampling error, the timeliness and the coherence.

Table 2. Quality aspects for stages of the statistical process

	Relevance	Sampling error	Non sampling error			Timeliness	Coherence
			Frame	Non response	Measurement		
Formulation user needs	+						+
Translation output specification	+						+
Using secondary sources	+				+	+	+
Sampling design / choice of sampling frame		+	+				
Questionnaire design					+		+
Drawing the sample							
Data collection		+		+	+	+	+
Follow up procedures		+		+		+	+
Data entry					+	+	+
Editing & imputation					+	+	+
Estimation procedure		+	+	+	+		+
Statistical integration	+	+	+	+	+	+	+
Analysis							
Tabulation & analysis results	+						
Disclosure control	+						

+ = activity influencing quality

Note that the use of external sources can be threefold. In the first phase secondary sources are used to complete a sample survey for its cut off or to shorten the questionnaire, in the second phase the sampling frame is used to develop an efficient sampling design, and in the third phase secondary sources are used to provide auxiliary information in the estimation procedure.

The relevance of a statistical product will depend on the discrepancy between the realized statistical output (tabulation and analysis results) and the intended statistical output. The intended statistical output is the result of a description of the users need. The discrepancy may be caused by 1) misinterpretation of the user needs, 2) shortcomings in the translation of the intended output into a formulation of the survey objectives, including the formulation of the output specifications, 3) discrepancy between the target population and the sampled population due to cut off, 4) bad choice of tabulation, and 5) disclosure control through which the tabulation and analysis results are less detailed than intended. By the use of secondary sources and statistical integration the gap between the realized and intended statistical output can be reduced.

The sampling error related to the gross sample depends on the sampling strategy, i.e. sampling design, gross sample size and estimation technique. The choice of the sampling design is based on the level of registered units in the sampling frame and on possibly additional information in the sampling frame. The choice of the estimation procedure depends on the sampling design and the availability of auxiliary information in external sources. Given a sampling strategy the sampling error related to the net sample is a function of the net sample size, design effect, and explanatory power of the auxiliary information in the estimation procedure. The difference between gross sample size and net sample size is caused by non response. The size of non response strongly depends on follow up procedures (or more generally field work organization) and data collection methods. If the statistical product is the outcome of statistically integrated (independent) surveys, then the sampling error of the (pooled) estimate depends on the separate sampling errors of the original estimates obtained by the surveys involved.

Given the formulation of the user needs and the output specification, non sampling errors may occur in almost every following activity of the statistical production process. Following Lessler and Kalsbeek (1992) we divide non sampling errors in frame errors, non response errors and measurement errors. Frame errors are caused by imperfections in the sampling frame, such as overcoverage, undercoverage and misclassification. Note that frame imperfections affecting cut-off and misclassification of statistical units are included in this type of error. Non response errors may arise from the failure to obtain response from all sample units. In general the response rate will depend on the data collection method. Response rates can be increased by means of follow up procedures. A measurement error occurs when an improper value is associated with the sampled unit. Measurement errors are caused by using secondary sources with slightly different definitions, item non response, proxy interviews, misinterpretations of questions, interviewer effects and errors in coding and typing. Note that item non response is considered as a measurement error rather than a non response error. Editing and imputation are often used techniques to handle measurement errors. All types of non sampling error may cause biased estimates. Methods to correct for bias are the use of external sources in the estimation procedure (re-weighting) and statistical integration.

All activities contribute to the lack of timeliness to a greater or lesser extent. However, the contribution to the timeliness of a number of activities can be considered fixed or negligible. These activities are: formulation of user needs, translation of output specification, defining the sampling frame, design of the sample and questionnaire, drawing the sample, estimation procedures, tabulation and analysis of results, and disclosure control. The remaining activities may influence the timeliness considerably. The use of secondary sources may have an impact on the timeliness, on the one hand due to the dependence on the register accountant, and on the other hand due to a reduction of the field work. The choice of the data collection method affects the timeliness greatly. For example, electronic retrieving data directly from respondents book keeping records (EDI) is much faster than data collection by mail. Follow up procedures will influence the timeliness in a straightforward way: more follow up procedures imply an increase in production time. Data entry, imputation or editing can be carried out by hand or automatically. If they are carried out by hand then the contribution to the timeliness is determined by the net sample size and if they are carried out automatically the contribution is negligible. Finally, statistical integration may lengthen the production time, due to the dependencies of the delivery of the results of other surveys/sources by means of which the comparisons are made.

Coherence with respect to definitions and classifications is mainly a result of a careful definition of the user needs and the corresponding translation into the intended output specifications. Despite the coherence with respect to definitions and classifications, different surveys may still produce mutually inconsistent outcomes, mainly due to non sampling errors. Use of registers (external sources) may disturb the coherence, because of discrepancies between target variables and registered variables or coordinated classifications and registered classifications. Different questions or routing in questionnaires concerning the same (target) variables may cause inconsistent outcomes, and so on. In the data processing phase inconsistencies can be removed or prevented. Imputation and editing may prevent inconsistent scores within records, predetermined consistency requirements in outcomes can be realized by specific weighting techniques, and remaining inconsistencies in outcomes can be removed subsequently by statistical integration.

5. The activity-costs matrix

When making a statistical product there are costs involved. For statistical agencies these costs concern personal and material costs. However, by filling in questionnaires, businesses take a substantial share in the production process of statistics (Willeboordse, 1995). Response burden can be defined as the total costs made by businesses to fill in questionnaires. As a consequence of non response, Willeboordse (1995) distinguishes between gross and net response burden. Gross response burden is defined as the total costs, which would have been made by businesses in filling in questionnaires if all sampled business respond. The total costs actually made by the responding businesses is called net response burden. So, besides the costs related to statistical agencies one should also incorporate costs (response burden), which are made by the businesses in order to obtain the total costs of a statistical product.

In order to improve a specific quality aspect, some activities have to be adjusted. In Table 2 these activities are marked for the quality aspect relevance, sampling error, non sampling error, timeliness and coherence. Often, an adjustment of a specific activity implies a change of the total costs. In Sections 5.1 and 5.2 it is discussed which activities cause a change of the total costs in order to achieve specific quality requirements with respect to the sampling error and the non sampling error respectively.

5.1. Costs with respect to the sampling error

Suppose that given a specific statistical production process the sampling error of most estimates is too high. According to Table 2, the sampling error is affected by the following activities: development of sampling design (including choice of sampling frame and choice of gross sample size), data collection, follow up procedures, estimation procedure (including the use of external sources), and statistical integration. The sampling error can be decreased by an increase of net sample size, by using a more efficient sampling design or by using more auxiliary information in the weighting procedure. For simplicity, it is assumed that only an increase of net sample size will change the total costs. Now, the net sample size can be increased by an increase of gross sample size and/or by carrying out more intensively follow up procedures. An increase of the gross sample size will increase on the one hand the costs made by the statistical agency concerning the data collection and on the other hand the gross response burden. Furthermore, an increase of gross sample size will result in an enlarged net sample size, which increases the net response burden. Carrying out more intensively follow up procedures will increase the costs made by the statistical agencies straightforwardly. It results in an increased net sample size, and therefore also in an increase of net response burden. Finally, an increased net sample size not only results in an increase of the net response burden, but also in increased costs made by statistical agencies with respect to the data entry and editing & imputation. In Table 3 it is schematically shown which activities cause a change of the total costs in order to achieve specific quality requirements with respect to the sampling error.

5.2. Costs with respect to the non sampling error

Deficiencies in the sample frame can be traced and eliminated by a more intensive quality control. Besides the usual maintenance, the quality of the frame may be improved by the use of additional sources such as external registrations, the survey samples drawn form the frame and samples specifically drawn for quality control. It is needless to say that more quality control is quite expensive for a statistical agency. Non response can be reduced by using a more suitable data collection method in combination with follow up procedures. This may affect the costs of the statistical agency and the net response burden. By applying quality control with respect to the use of secondary sources, questionnaire design, data collection method and data entry the magnitude of measurement errors can be minimized. After the data entry the raw data can be checked and if necessary corrected by editing and imputation methods.

In all these activities costs for the statistical agency are involved. With respect to editing and imputation the net response burden will be influenced too. Finally, statistical agencies incur costs to reduce potential bias caused by all types of non sampling error by using advanced estimation techniques and statistical integration. Per type of non sampling error it is schematically shown in Table 4 which activities cause a change of the total costs in order to achieve specific quality requirements. We will continue the discussion on costs in Section 9.

Table 3. Costs with respect to the sampling error related to the activities

	Costs made by statistical agencies	Gross response burden	Net response burden
Formulation user needs			
Translation output specification			
Using secondary sources			
Sampling design / choice sampling frame			
Questionnaire design			
Drawing the sample			
Data collection	+	+	+
Follow up procedures	+		+
Data entry	+		
Editing & imputation	+		+
Estimation procedures, using external sources			
Statistical integration			
Analysis			
Tabulation & analysis results			
Disclosure control			

+ = non negligible influence on costs

Table 4. Costs with respect to the non sampling error related to the activities

	Frame error	Non response error		Measurement error	
	Costs	Costs	Net response burden	Costs	Net response burden
Formulation user needs					
Translation output specification					
Using secondary sources				+	
Sampling design / choice sampling frame	+				
Questionnaire design				+	
Drawing the sample					
Data collection		+	+	+	
Follow up procedures		+	+		
Data entry				+	
Editing & imputation				+	+
Estimation procedures, using external sources	+	+		+	
Statistical integration	+	+		+	
Analysis					
Tabulation & analysis results					
Disclosure control					

+ = non negligible influence on costs

Part B. The model

6. Overview

In Section 7 we describe an instrument by means of which a survey manager may judge the quality of the complete statistical production process, that is, from the use and maintenance of the sampling frame to the development of the estimation procedure. This instrument should give the survey manager a first impression of the attempts that have been made to prevent, control, and correct for survey errors that might arise at nearly every stage of the production process (see Part A, Section 1 for a subdivision of the total error). The idea is to employ various qualitative indicators corresponding to the efforts made to prevent, control, and correct for a number of error components, and to summarize the ordinal scores of each quality indicator into a 'score matrix' (see Table 5). The higher a specific score, the more efforts are made (on average) to prevent, control, or correct for a particular error. The rationale behind this score matrix is that the probability of large errors are probably less in case of high scores. It should be considered as a means to judge the quality of the statistical production process, although there is no fixed (measurable) relation to the quality of the statistical product, i.e. the total survey error.

In Section 8 we elaborate on the activities which are related to the sampling error. We discuss two quantitative quality indicators; one corresponding to data collection and follow up procedure, and the other to sampling design and estimation procedure. As can be seen from Table 2, these activities not only influence the sampling error, but they also affect the non sampling error, in particular the non sampling error due to non response. This implies that both quality indicators provide (quantitative) information about the sampling error as well as (a part of) the bias, namely the bias due to non response. Given the total survey costs, it holds for most statistical production processes that there is a trade off between reducing sampling error and reducing bias due to non response. This also follows from the cost function, which will be discussed in Section 9. In the ideal situation one should strive for an optimal survey strategy minimizing the mean squared error, subject to the cost restrictions. However, because it is very difficult to measure the bias (contrary to measuring the sampling error) and hence the mean squared error, such a strategy is often not feasible in practice. We therefore suggest a less ambitious strategy which relates sampling error to non response instead of bias. This strategy is discussed in Sections 10 and 11 and is implemented in a software package (see Part C). It constitutes the basis of our model. Now, this model can be extended in several ways. In the remaining sections of this part it is shown how e.g. the net response burden and the frame errors can be incorporated. However, these ideas are not fully elaborated.

In the literature, sophisticated survey error models, such as variance decomposition models, have been developed in an attempt to quantify the seriousness of the error components (e.g. Lessler and Kalsbeek, 1992). We recognize the importance of such models, because they provide an (objective) measure of the relative contribution of each error to the total survey error taking into account the interrelationships among the sources of errors. However, these models require experimental methods, such as repeated measurements or interpenetrating samples, for quantifying the components. In the absence of such experiments, one should settle for a less ambitious goal.

7. Quality indicators with respect to the non sampling errors

In this section we concentrate on the non sampling error. We distinguish between frame errors, measurement errors and non response errors (see Part A, Section 4). Non sampling errors may occur in almost every activity of the survey process. Each activity can therefore be considered as a potential instrument for quality control. The process characteristics, the actual realization of the survey activities, give an indication of what has been done to minimize each error. A well known measure for non sampling errors is the bias. However, components of the bias are hard to quantify in practice. For this reason the non sampling errors will be evaluated on the basis of the efforts made to reduce these errors. That is, a number of relatively simple quality indicators will be constructed by assigning scores to the process characteristics influencing the non sampling errors. A questionnaire was designed to collect the process information needed from the Member States (Appendix D).

The magnitude of the non sampling errors is assumed to depend on the efforts made to prevent, control and correct for these errors. Given these three elements of quality control, Table 5 shows eleven quality indicators corresponding to the activities distinguished in the survey process. Five indicators relate to frame errors, four indicators relate to measurement errors, and two indicators are concerned with non response. Note that we see item non response as a measurement error and not as a non response error (see Section 4). The quality indicators cover the information that EUROSTAT expects the Member States to provide about non sampling errors as proposed by the working group "Assessment of the quality of business statistics" (see EUROSTAT, 1996a). In fact, most of the process characteristics underlying the quality indicators will be asked by EUROSTAT too. Our questionnaire therefore serves as a kind of test case for the Member States.

The question is how to translate the process information obtained from the questionnaire into scores on the quality indicators. A computer program was developed to enter the data and to aid the translation process (see Part C). First, the answers of the relevant questions are coded into a limited number of pre-defined answer categories (see appendix D). As the answers are often descriptive in nature, the coding may not always be straightforward and could be subject to interpretative differences. In appendix E we give some guidelines that can be helpful when dealing with this problem. We like to point out that it is not our objective to specify a normative framework for the coding procedure. The pre-defined answer categories form the input of the program. Next, a numerical score is assigned to each answer category (see fifth column 'Range of scores', Table 5). In case a question is left open or is answered as 'not known' a nil score is given.

Table 5. Quality indicators grouped by activity and process characteristics

Activity	Process characteristics	Question	Range of scores a)	Quality indicator	Range of scores b)
Maintenance of register	• inspection of births • size of quality survey	• Q6 • Q11a	• [0, 2] • [0 ,2]	checking units	0 $\rightarrow 1w_1$ 1 $\rightarrow 2w_1$ 2 $\rightarrow 3w_1$ 3,4 $\rightarrow 4w_1$
Maintenance of register	• correction for misclassifications • correction for deaths • correction for duplicates • size of quality survey	• Q8 • Q9 • Q10 • Q11a	• [0, 2] • [0, 2] • [0, 2] • [0, 2]	correcting units	0,1 $\rightarrow 1w_2$ 2,3 $\rightarrow 2w_2$ 4,5 $\rightarrow 3w_2$ 6,7,8 $\rightarrow 4w_2$
Maintenance of register	• timeliness registration of births • frequency of quality survey	• Q7 • Q11b	• [0, 2] • [0, 2]	updating units	0 $\rightarrow 1w_3$ 1,2 $\rightarrow 2w_3$ 3 $\rightarrow 3w_3$ 4 $\rightarrow 4w_3$
Use of sampling frame	• missing categories • other types of undercoverage	• Q12a • Q12b	• [0, 2] • [0, 2]	undercoverage	0,1 $\rightarrow 1w_4$ 2 $\rightarrow 2w_4$ 3 $\rightarrow 3w_4$ 4 $\rightarrow 4w_4$
Use of sampling frame	• dead or misclassified units • detection of overcoverage	• Q12c • Q13	• [0, 1] • [0, 2]	overcoverage	0 $\rightarrow 1w_5$ 1 $\rightarrow 2w_5$ 2 $\rightarrow 3w_5$ 3 $\rightarrow 4w_5$
Questionnaire design	• participation of respondents • support in completing questionnaire	• Q19 • Q20	• [0, 2] • [0, 2]	preventing measurement errors (incl. item non response)	0 $\rightarrow 1w_6$ 1 $\rightarrow 2w_6$ 2 $\rightarrow 3w_6$ 3,4 $\rightarrow 4w_6$
Data collection and follow up procedures	• non response sensitivity of data collection method • intensity follow up procedure	• Q23 • Q24	• [0, 2] • [0, 2]	preventing unit non response	0 $\rightarrow 1w_7$ 1 $\rightarrow 2w_7$ 2 $\rightarrow 3w_7$ 3,4 $\rightarrow 4w_7$
Editing & imputation	• intensity of checks on records • detection of records requiring manual review	• Q36 • Q37	• [0, 2] • [0, 2]	checking measurement errors	0 $\rightarrow 1w_8$ 1,2 $\rightarrow 2w_8$ 3 $\rightarrow 3w_8$ 4 $\rightarrow 4w_8$

Table 5. Quality indicators grouped by activity and process characteristics (continued)

Activity	Process characteristics	Question	Range of scores a)	Quality indicator	Range of scores b)
Imputation	• qualification of imputation method • extent of imputations • recognizing imputed items • evaluation impact of imputations on estimates	• Q31 • Q33 • Q32 • Q34	• [0, 2] • [0, 1] • [0, 1] • [0, 1]	correcting measurement errors I (incl. item non response)	0,1 $\rightarrow 1w_9$ 2 $\rightarrow 2w_9$ 3 $\rightarrow 3w_9$ 4,5 $\rightarrow 4w_9$
Editing	• qualification of edit system • recognizing edits • evaluation impact of editing on estimates	• Q35, Q38 • Q39 • Q40	• [0, 2] • [0, 1] • [0, 1]	correcting measurement errors II	0,1 $\rightarrow 1w_{10}$ 2 $\rightarrow 2w_{10}$ 3 $\rightarrow 3w_{10}$ 4 $\rightarrow 4w_{10}$
Estimation	• methods to correct for non response • oversampling for non response • recognizing imputed units • evaluating impact of weight adjustment/imputation on estimates	• Q28 • Q17 • Q29 • Q30	• [0, 2] • [0, 1] • [0, 1] • [0, 1]	correcting unit non response	0 $\rightarrow 1w_{11}$ 1 $\rightarrow 2w_{11}$ 2,3 $\rightarrow 3w_{11}$ 4,5 $\rightarrow 4w_{11}$

a) ordinal scale running from a low score (0) to a high score (max. 2)

b) ordinal 4-points scale running from poor ($1w_i$) to high quality ($4w_i$) derived by summing the scores on the underlying process characteristics, where w_i denotes the weight assigned to quality indicator i (i=1,2,...,11)

The scores on the quality indicators are subsequently obtained by simply adding up the scores on the underlying process characteristics. Finally, these scores are transformed into a 4-points scale, where '$1w_i$' indicates poor quality and '$4w_i$' indicates a high level of quality (see Table 5). By giving each quality indicator a weight w_i the survey manager can set priorities. The weights are calculated as $w_i = r_i \Big/ \sum_{i=1}^{11} r_i$, where r_i denotes the 'level of importance' that is assigned to quality indicator *i*: low ($r_i = 1$), medium ($r_i = 2$) or high ($r_i = 3$). For example, if two indicators are labeled 'high' and the remaining indicators are labeled 'medium', the indicators get the weights 1/8 = 3/(2×3 + 9×2) and 1/12 = 2/(2×3 + 9×2), respectively. Further aggregation is possible by clustering the quality indicators into groups, for example 'frame errors', 'measurement errors' and 'non response', and taking the weighted sum of the scores within each cluster as the group score.

The quality indicators can be considered as indirect measures of the potential non sampling errors in the estimates. The indicators do not measure the bias, but show how adequate the instruments (i.e., process activities) are in controlling the quality with respect to the non sampling errors. Essential for quality control is that one has a clear understanding of the survey process. If all questions about the process characteristics are answered completely, Member States show that they have full control over the survey process. Even if some scores on the process characteristics indicate that some instruments may not be very effective in reducing non sampling errors, the process can still be under control. On the other hand, if the information on process characteristics is incomplete then this could indicate that one has insufficient knowledge about some aspects of the process, and hence that the process is not fully under control. So, in our opinion it is important to make a distinction between the extent and the effectiveness of quality control. The total number of questions left open or answered as ‘not known’ is taken as a measure of process control. The higher the score, the less control one has over the different stages of the survey process. The scores on the eleven quality indicators in Table 5 provide a measure of the effectiveness of quality control.

8. Quantitative quality indicators with respect to the sampling error

As shown in Part A, Section 4 (Table 2), the sampling error is influenced by several activities and characteristics of the statistical process. The sampling error depends on the sampling design, the net sample size, and the estimation technique (including the use of auxiliary information). The net sample size, i.e., the actual response, is equal to the gross sample size minus the non response (for the moment we assume that there are no frame errors). The (non) response in turn depends on the follow up procedure and the data collection method. In order to quantify the sampling error a model has been developed describing all these relations. The sampling design may be less effective than intended because of frame errors. The relation between sampling error and frame error will be discussed separately in Section 12.

In the following we consider stratified simple random sampling. The population of N units is stratified into H strata of sizes N_h and in h-th stratum a simple random sample of size n_h^{gross} is drawn. However, due to non response, only n_h^{net} units are actually observed. These net sample sizes will be the key factor of our model. Given the gross sample size, the net samples size can be increased by carrying out follow up procedures. A follow up procedure may consist of several attempts to increase the response and the success rate may differ for every attempt. Assuming that all units who did not respond at the previous attempt(s) are contacted again at the next attempt we get for each stratum the following response model:

1st attempt: $\lambda_{h1} n_h^{gross}$ units responding,

2nd attempt: $\lambda_{h2}(1-\lambda_{h1})n_h^{gross}$ units responding,

...

…

kth attempt: $\lambda_{hk}(1-\lambda_{h(k-1)})\ldots(1-\lambda_{h1})n_h^{gross}$ units responding,

where λ_{hi}, i=1,..., k_h, are the success rates for the i-th attempt in stratum h to contact the respondents. Figure 2 further illustrates the response model. Every attempt to contact units will increase the total response and hence reduce the non response. The success rates λ_{hi} determine to what extent and how fast the total response will be increased.

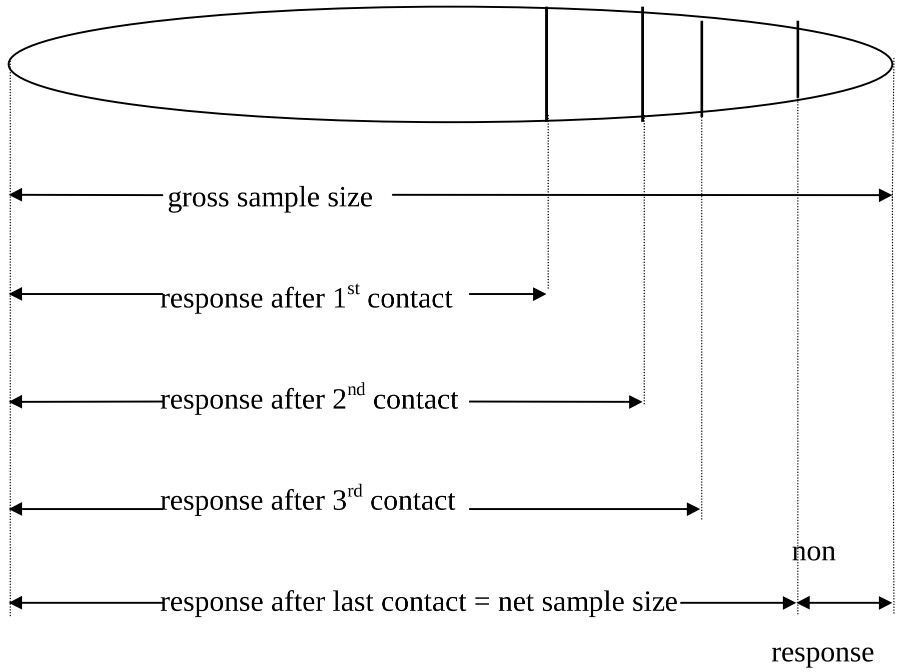


Fig. 2. Subdivision of the sample in response and non response

The net sample size equals the total response obtained at the end of the follow up procedure. According to our model we get, by summing the response over all attempts, the following net sample size:

$$n_h^{net} = \left[1 - \prod_{i=1}^{k_h} (1 - \lambda_{hi})\right] n_h^{gross} = L_h n_h^{gross}, \qquad (1)$$

where k_h is the total number of attempts to contact units in stratum *h*. Here,

$$L_h = \left[1 - \prod_{i=1}^{k_k} (1 - \lambda_{hi})\right] \qquad (2)$$

is the modeled response fraction in the *h*-th stratum. Note that after data collection (2) can be estimated by the realized response fraction in stratum *h*. The modeled total number of units contacted (successfully or unsuccessfully) in the *h*-th stratum is equal to

$$n_h^{contact} = \left[1 + \sum_{j=1}^{k_h - 1} \prod_{i=1}^{j} (1 - \lambda_{hi})\right] n_h^{gross} = M_h n_h^{gross}, \qquad (3)$$

where

$$M_h = \left[1 + \sum_{j=1}^{k_h - 1} \prod_{i=1}^{j} (1 - \lambda_{hi})\right] \qquad (4)$$

can be considered as the modeled mean contact rate in stratum h: $M_h = n_h^{contact} \big/ n_h^{gross}$. Again, after data collection (4) can be estimated by the realized mean contact rate. Note that the gross sample sizes n_h^{gross} and the number of contact attempts k_h are instruments of the survey process controlled by the survey manager. The success rates λ_{hi} on the other hand are parameters of the model, which need to be determined from the data. These fractions will to a large extent depend on the method of data collection and the fieldwork organization. In most practical situations we will have $\lambda_{h1} \geq \lambda_{h2} \geq \ldots \geq \lambda_{hk_h}$. However, if $\lambda_{h1} = \lambda_{h2} = \ldots = \lambda_{hk_h} = \lambda_h$ then there is a simple relationship between M_h and L_h, namely $L_h = \lambda_h M_h$.

The distribution of the gross sample size over the strata is determined by the gross allocation scheme. Let $\xi_h^{gross} = n_h^{gross} \big/ n^{gross}$ denote the particular gross allocation employed for stratum h. Two well-known allocation scheme's are proportional allocation and the so-called Neyman allocation:

$$\xi_h^{prop} = \frac{N_h}{N} \quad \text{and} \quad \xi_h^{neyman} = \frac{N_h S_{yh}}{\sum_h N_h S_{yh}} \text{, where}$$

S_{yh} is the population standard deviation in stratum h: $S_{yh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \overline{Y}_h)^2$. A third allocation scheme, namely optimum allocation, will be discussed in Section 8. This optimum allocation scheme not only takes into account a cost function, which will be discussed in the next section, but also the modeled response fractions given by (2).

On the basis of the gross allocation scheme we define the following (quantitative) quality indicator for the total response fraction:

$$L = \frac{n^{net}}{n^{gross}} = \frac{\sum_{h=1}^{H} n_h^{net}}{n^{gross}} = \frac{\sum_{h=1}^{H} n_h^{gross} L_h}{n^{gross}} = \sum_{h=1}^{H} \xi_h^{gross} L_h . \qquad (5)$$

This response rate can be interpreted as a weighted mean of the stratum response fractions L_h with the gross allocations ξ_h^{gross} as weights. Clearly, the total response fraction as defined by (5) depends on the gross allocation scheme, which is chosen by the survey designer.

As already said, non response has two main consequences, namely 1) a reduction of the net sample size, and hence an increase in variance, and 2) an increase in bias. An indicator of the former is given by an estimate of the variance, see (9), whereas an indicator of the latter is given by (5). In general, (5) may be a poor measure of bias because the exact relation between the response fraction and the bias is not known. Instead we have to rely on the following rules of thumb which are generally known for stratified designs. The bias tends to be smaller when 1) the response rates L_h are higher and 2) the stratum population variances S_h^2 are smaller, see Bethlehem and Kersten (1985, p. 290). These rules lead us to the following definition:

$$L(y) = \left(\sum_{h=1}^{H} N_h S_{yh} L_h \right) \Bigg/ \left(\sum_{h=1}^{H} N_h S_{yh} \right). \qquad (6)$$

The total response fraction is defined as a weighted mean of the stratum response fractions with weights corresponding to the Neyman allocation. A relatively large weight is assigned to a stratum if the stratum is larger or the stratum is more variable internally. We note that the total response fraction as defined by (6) depends on the target variable *y*, as it should be, because the seriousness of the bias due to non response may vary from variable to variable.

In order to define a (quantitative) quality indicator for the sampling error, we assume that each stratum total is estimated by a regression estimator. Let *y* be a scalar target variable and *x* a *p*-vector valued auxiliary variable that is correlated with *y*. Using the separated general regression estimator, the estimator for the population total of *y* is given by

$$\hat{Y}_r = \sum_{h=1}^{H} N_h \left[\bar{y}_h + \hat{B}_h^t (\bar{X}_h - \bar{x}_h) \right], \tag{7}$$

where $\bar{y}_h$ is the (net) sample mean of *y* for stratum *h*, $\bar{x}_h$ is the (net) sample mean of x for stratum *h*, $\bar{X}_h$ is the population mean of *x* for stratum *h*, and

$$\hat{B}_h = \left(\sum_{i=1}^{n_h^{net}} \gamma_{hi} x_{hi} x_{hi}^t \right)^{-1} \left(\sum_{i=1}^{n_h^{net}} \gamma_{hi} x_{hi} y_{hi} \right) \tag{8}$$

with $\gamma_{hi} > 0$. It is required that all γ_{hi} are known. Särndal et al. (1992) suggested taking $\gamma_{hi} = \sigma_{hi}^{-2}$, where σ_{hi}^2 can be interpreted as the variance of independent random variables y_{hi}^* defined in a superpopulation model of which the y_{hi} are supposed to be the outcomes. Note that each term in (7) is a general regression estimator for the stratum population total Y_h. Note further that the ratio estimator, the ordinary regression estimator and post-stratification are all special cases of the general regression estimator (see e.g. Särndal et al., 1992, Chap. 7). Let $e_{hi} = y_{hi} - B_h^t x_{hi}$, then the variance of the separated regression estimator for the population total can be approximated by (Särndal et al., 1992, Chap. 6):

$$Var(\hat{Y}_r) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h^{net}} - \frac{1}{N_h} \right) S_{eh}^2 , \tag{9}$$

with

$$S_{eh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (e_{hi} - \bar{E}_h)^2 \quad \text{and} \quad B_h = \left(\sum_{i=1}^{N_h} \gamma_{hi} x_{hi} x_{hi}^t \right)^{-1} \left(\sum_{i=1}^{N_h} \gamma_{hi} x_{hi} y_{hi} \right).$$

If there exists constant *p*-vectors c_h such that $c_h^t x_{hi} = \lambda_{hi}^{-1}$ for all $i = 1, \ldots, N_h$ then $\bar{E}_h = 0$ for all $h = 1, \ldots, H$ (compare result 6.5.1 of Särndal et al. 1992). If in addition $\gamma_{hi} = 1$ we may rewrite S_{eh}^2 in terms of unexplained variation. Let ρ_h^2 denote the (multiple) correlation coefficient between *y* and *x* in stratum *h*, i.e.

$$\rho_h^2 = \frac{B_h^t \left(\sum_{i=1}^{N_h} x_{hi} x_{hi}^t \right) B_h - N_h \overline{Y}_h^2}{\sum_{i=1}^{N_h} y_{hi}^2 - N_h \overline{Y}_h^2}, \qquad (10)$$

then it is easy to show that $S_{eh}^2 = S_{yh}^2 (1-\rho_h^2)$. The decrease of the variance, due to the use of auxiliary information, is now easily measured by means of the multiple correlation coefficients. Therefore, we will associate the multiple correlation coefficient with the use of auxiliary information. We note that all parameters required for estimating S_{eh}^2 and/or ρ_h^2 are based on the net sample size.

The use of auxiliary information is a well known method to improve quality. It is often possible to gather additional information for the sample elements without much extra effort or costs. In business statistics the number of employees is often used as an auxiliary variable. It is generally known that auxiliary information can be used to 1) reduce sampling variance, 2) correct for non response bias, and 3) make the estimates consistent with other population parameters. Note that for variables to be used as auxiliary information the population totals have to be known. For convenience it was assumed that this auxiliary information is incorporated in the estimation procedure by means of the separated regression estimator (7). For large samples the variance of this estimator is given by formula (9). As can be seen from this formula, the use of auxiliary information influences the variance rather straightforwardly, i.e. it reduces the stratum variances by a factor $(1-\rho_h^2)$. If this factor varies between strata then the use of auxiliary information also influences the Neyman allocation. Hence, an improved version of (6) is

$$L(y) = \frac{\sum_{h=1}^{H} N_h S_{yh} \sqrt{(1-\rho_h^2)} L_h}{\sum_{h=1}^{H} N_h S_{yh} \sqrt{(1-\rho_h^2)}}. \qquad (11)$$

Obviously, (11) not only depends on the response fractions L_h and the stratum variances S_{yh}^2, but also on the correlation coefficients ρ_h^2. The higher the squared correlation coefficient in a specific stratum, the lower the stratum contribution with respect to the total response fraction. This corresponds to the fact that the use of auxiliary information may reduce the bias. It should therefore be incorporated into the measure related to non response.

9. Cost functions with respect to sampling error

With respect to the costs of business surveys, statistical agencies are asked to provide the following information: the total personnel costs and the total material costs of the survey, the distribution of personnel over the survey activities, and some information relating to the costs made by businesses in completing the survey. These cost characteristics form the input of the model. In addition, information about the survey methods that are used is required. First of all, the costs of a survey depend on the process characteristics. The use of different methods will usually lead to differences in costs as well. In Section 2 we have discussed seven activities, of which only four are related to the total costs with respect to the sampling error (see Table 3). These activities are data collection, follow up procedures, data entry, and editing and imputation. In this section, we will concentrate on these activities.

Again, we assume that the survey design is stratified simple random. For each stratum we distinguish between n_h^{gross}, $n_h^{contact}$, and n_h^{net}. These numbers are related through the number of follow up procedures k_h and the success rate at each trial λ_{hi}, as specified by (1) and (3). In general, the costs will depend on the number of units included in a specific activity. For example, the more questionnaires are collected and processed, or the more units are contacted in the follow up procedure, the higher the costs will be. The costs of a specific activity *m* in stratum *h* can be described as a function of the method used, T_h^m, and the number of units included in the activity:

$$C_h^m = C_h^m(T_h^m, n_h^{\cdots}), \tag{12}$$

where $n_h^{\cdots}$ can be equal to n_h^{gross}, $n_h^{contact}$, and n_h^{net}, depending on the activity under consideration. As stated in Section 7, the net sample size and the number of units contacted depend on the total number of attempts to contact units, the response fractions and the gross sample size. Hence, the costs can also be expressed as a function of k_h and λ_{hi} and n_h^{gross}. These relations will be further elaborated below. In the following we will assume that

- each cost function (12) has two components: fixed costs and variable costs,
- the variable costs are proportionally related to the number of units included in the activity,
- for each activity no more than one method is used (if more than one method is used for an activity, only the most important method will be considered).

The last assumption implies that the costs of each activity are equal to the costs of the (main) method used in each activity. The total costs made by the statistical agency per stratum, C_h, are equal to the sum of the costs of the different activities distinguished in the survey process, i.e.,

$$C_h = \sum_m C_h^m \,. \tag{13}$$

For reasons of completeness the general cost function (12) is worked out below for all seven activities mentioned in Section 2.

For the activity 'use of the sampling frame' (*m* = 1) two methods are distinguished, namely maintenance of the sampling frame without quality inspection, or maintenance of the sampling frame including quality inspection. In each stratum one can choose either one. Once the choices are made, the costs can be considered fixed in each stratum:

$$C_h^1 = c_h^1. \tag{14}$$

For the activity 'questionnaire design' (*m* = 2) we distinguish between three methods, which can be classified as: evaluation methods with no respondents involved, evaluation methods with few respondents involved, or evaluation methods with many respondents involved. When different questionnaires are used for the different strata, the evaluation methods may differ for the strata as well. Once the methods are selected, the costs can be considered fixed in each stratum:

$$C_h^2 = c_h^2 . \tag{15}$$

For the activity ‘data collection’ (m = 3) we distinguish between four methods: mail, computer assisted telephone interviewing (CATI), computer assisted personal interviewing (CAPI), or electronic data interchange (EDI). In principle, different data collection modes can be used for different strata. Once the choices are made, the costs of the data collection depend on the gross sample size:

$$C_h^3 = c_h^3 + a_h^3 n_h^{gross} , \tag{16}$$

where a_h^3 are the costs per unit (at the first trial). Note that these costs per unit depend on the kind of data collection method. For the activity ‘follow up’ (m = 4) we assume only variable costs:

$$C_h^4 = a_h^4 n_h^{follow} , \tag{17}$$

where n_h^{follow} is the total number of contacted units in the follow up procedure: $n_h^{follow} = n_h^{contact} - n_h^{gross}$.

We distinguish between three follow up methods, namely : follow up by letter, follow up by telephone, or follow up by field workers. We note that the costs per unit in the “follow up” may differ from the costs per unit in the “data collection”, for example because different modes are used. For the activity ‘data entry’ (m = 5) the cost function is

$$C_h^5 = c_h^5 + a_h^5 n_h^{net} , \tag{18}$$

where the costs per unit a_h^5 depend on whether the data entry is done by hand, or automatically. For the activity ‘editing and imputation’ (m = 6) the cost function is

$$C_h^6 = c_h^6 + a_h^6 n_h^{net} . \tag{19}$$

The costs a_h^6 per unit depend on the method used for editing and imputation: manual, semi-automatic, or full automatic. Finally, for the activity ‘estimation’ (m = 7) and all other survey activities we only consider fixed costs:

$$C_h^7 = c_h^7 . \tag{20}$$

Summing the results and making use of (1) to (4), the total costs per stratum can be expressed as

$$C_h = c_h + a_h n_h^{net} \tag{21}$$

where

$$c_h = c_h^1 + c_h^2 + c_h^3 + c_h^5 + c_h^6 + c_h^7 \tag{22}$$

are the total fixed costs of stratum h and

$$a_h = \frac{a_h^3}{L_h} + \frac{a_h^4 (M_h - 1)}{L_h} + (a_h^5 + a_h^6) \tag{23}$$

are the variable costs per (net) sampling unit. If $a_h^3 = a_h^4$, i.e. the costs per unit in the "follow up" equals the costs per unit in the "data collection", then (23) can be simplified to

$$a_h = \frac{a_h^4 M_h}{L_h} + (a_h^5 + a_h^6)\,. \qquad (23a)$$

Note that a_h given by (23a) is a function of the total number of contact attempts k_h through M_h / L_h. It is shown in appendix A that $M_h(k_h)/L_h(k_h) \geq M_h(k_h - 1)/L_h(k_h - 1)$ if $\lambda_{h1} \geq \lambda_{h2} \geq \ldots \geq \lambda_{hk_h}$. So, in most practical situations a_h is non-decreasing in k_h. In other words, the variable costs per (net) sampling unit will generally increase as the follow up procedure is intensified. If $\lambda_{h1} = \lambda_{h2} = \ldots = \lambda_{hk_h} = \lambda_h$, so $L_h = \lambda_h M_h$, then (23) can be further simplified:

$$a_h = \frac{a_h^4}{\lambda_h} + (a_h^5 + a_h^6) \equiv a'_h\,. \qquad (23b)$$

In this particular case a_h is only dependent on the cost and data characteristics. Finally, the total costs can be expressed as

$$C = \sum_{h=1}^{H} C_h = c_0 + \sum_{h=1}^{H} a_h n_h^{net}\,. \qquad (24)$$

The costs are measured in local currency and will be converted to euro's for reasons of comparability. Differences in the cost structure can give some indication of the relative contribution of different survey activities to the total efficiency. However, we do not consider it very useful to compare the surveys of the Member States on the basis of absolute differences in costs.

An important issue that remains to be considered is to determine the cost parameters. These parameters are generally not known and have to be estimated from the process and cost information obtained from the Member States. We concentrate on the cost function which is given by (21). This cost function contains two unknown parameters, namely c_h given by (22) and a_h given by (23). These parameters can be estimated at two levels, depending on the amount of information available about the surveys costs. First, suppose that there is very little information, i.e., per stratum information is only available about the total fixed costs C_h^F and the total variable cost C_h^V. By definition we have

$$C_h = C_h^F + C_h^V \equiv (\bar{\gamma}_h^F + \bar{\gamma}_h^V) C_h\,,$$

where $\bar{\gamma}_h^F$ is the proportion of fixed costs and $\bar{\gamma}_h^V$ is the proportion of variable costs. These proportions are assumed to be known for the survey process in question. Rough estimates for c_h and a_h are now obtained by C_h^F and C_h^V / n_h^{gross} respectively. That is, the estimated costs per unit are equal to the average of the total variable costs.

A more refined approach is first to estimate all fixed cost components c_h^m and variable cost components a_h^m, and then to insert these estimates into (22) and (23) to obtain estimates for c_h and a_h respectively. Decompose the total costs with respect to activity *m* in stratum *h* into

$$C_h^m = F_h^m + V_h^m ,$$

where F_h^m are the fixed costs with respect to activity *m* and V_h^m the variable costs with respect to activity *m*. The fixed costs component c_h^m can be estimated by F_h^m and the costs per unit a_h^m by $V_h^m / n_h^{...}$. In principle, the costs of each activity and the method used can be derived directly from the questionnaire sent to the statistical agencies in the Member States. This approach needs separate estimates for c_h and a_h. If these costs estimates are not available for all strata then one may consider to collapse some strata, assuming that these parameters are equal for the collapsed strata.

The division into fixed and variable costs depends on the activity in question. Some activities have only fixed costs. If variable costs exist, the proportion of fixed and variable costs of the activity in question must be specified in advance. As information on fixed and variable costs is often not available, rough estimates will be used. We note that for the trade-off between costs and precision (see Section 11) only the variable costs are relevant.

10. Optimum allocation given the total budget restriction

10.1 Optimum allocation for a fixed follow up strategy

An important issue in stratified designs is the allocation scheme, i.e. the allocation of the sample over the strata. In this section we distinguish between the gross allocation scheme $\xi_h^{gross} = n_h^{gross} / n^{gross}$ and the net allocation scheme $\xi_h^{net} = n_h^{net} / n^{net}$. We concentrate on the net allocation scheme; the gross allocation scheme is easily obtained by the following relation:

$$\xi_h^{gross} = \frac{1}{L_h} n_h^{net} \Bigg/ \sum_{h=1}^{H} \frac{1}{L_h} n_h^{net} = \frac{1}{L_h} \xi_h^{net} \Bigg/ \sum_{h=1}^{H} \frac{1}{L_h} \xi_h^{net} . \qquad (25)$$

The problem is to find a net allocation scheme such that the variance given by (9) is minimized subject to the cost function given by (24). For fixed L_h and M_h, that is for fixed k_h, the optimum net allocation scheme is found to be (see e.g. Cochran, 1977, Chap. 5)

$$\xi_h^{net} = \frac{N_h S_h \sqrt{(1-\rho_h^2)} \Big/ \sqrt{a_h}}{\sum_{h=1}^{H} \left(N_h S_h \sqrt{(1-\rho_h^2)} \Big/ \sqrt{a_h} \right)} . \qquad (26)$$

Pre-multiplying the right and left hand side of (26) by $a_h n^{net}$ and summing over the strata we obtain

$$\sum_{h=1}^{H} a_h n_h^{net} = \frac{\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\sqrt{a_h}}{\sum_{h=1}^{H}\left(N_h S_h \sqrt{(1-\rho_h^2)}\Big/\sqrt{a_h}\right)} n^{net} . \tag{27}$$

Inserting (27) into (24) we obtain

$$C - c_0 = \frac{\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\sqrt{a_h}}{\sum_{h=1}^{H}\left(N_h S_h \sqrt{(1-\rho_h^2)}\Big/\sqrt{a_h}\right)} n^{net} \tag{28}$$

and by inserting (28) into (26) we find

$$n_h^{net} = (C - c_0)\frac{N_h S_h \sqrt{(1-\rho_h^2)}\Big/\sqrt{a_h}}{\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\sqrt{a_h}} . \tag{29}$$

Unless the cost function is given by (23b), these net stratum sizes are dependent on the follow up intensities $k_1, \ldots, k_H$. The corresponding minimum variance is obtained by inserting (29) into (9)

$$V_{\min} = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\sqrt{a_h}\right)^2}{C - c_0} - \sum_{h=1}^{H} N_h S_h^2 (1-\rho_h^2) . \tag{30}$$

Again, unless the cost function is given by (23b), this minimal variance is a function of the follow up intensities $k_1, \ldots, k_H$: $V_{\min} = V_{\min}(k_1, \ldots, k_H)$. Given these k_h's, the optimal net sample sizes can be obtained by the following choices of the gross sample sizes

$$n_h^{gross} = \frac{1}{L_h(k_h)} n_h^{net}(k_1, \ldots, k_H) , \tag{31}$$

where L_h and n_h^{net} are given by (2) and (29) respectively. Equation (31) is important because it relates the follow up intensities and the gross sample sizes for fixed costs *C*. In most practical situations we have $\lambda_{h1} \geq \lambda_{h2} \geq \ldots \geq \lambda_{hk_h}$ in which case a_h is non-decreasing in k_h (with $k_1, \ldots, k_{h-1}, k_{h+1}, \ldots, k_H$ fixed). It follows that the numerator of (29) is non-increasing while the denominator is non-decreasing in k_h. We conclude that (for fixed costs *C*) n_h^{net} is non-increasing in k_h. Furthermore, from (2) it is easily seen that L_h is non-decreasing in k_h. So, by (31) we also have that n_h^{gross} is non-increasing in k_h. The negative relation between n_h^{net} and k_h seems contradictory, but can be explained as follows. An increase in k_h by one has two effects on n_h^{net}, namely an increase in n_h^{net} by an amount of $\lambda_{hk}(1-\lambda_{h,k-1})\ldots(1-\lambda_{h1})n_h^{gross}$ and a decrease in n_h^{gross} due to the budget restrictions. Obviously, the first effect is completely offset by the second effect if

$\lambda_{h1} \geq \lambda_{h2} \geq \ldots \geq \lambda_{hk_h}$. We note that this relation between follow up intensity and net and gross sample sizes no longer holds in case of exhaustive or nearly exhaustive strata, see Section 10.3.

10.2 Balancing between precision and response fraction

According to Section 10.1 there is a trade off between the precision and response rate. In order to get some insight into this trade off we first discuss two special cases. These cases are mainly used as an illustration, although they are probably too restrictive to be of practical interest.

- $\lambda_{h1} = \lambda_{h2} = \ldots = \lambda_{hk_h} = \lambda_h$ with selective non-response, and
- $\lambda_{h1} \geq \lambda_{h2} \geq \ldots \geq \lambda_{hk_h}$ with non-selective non-response.

In the first case, the cost parameters a_h are independent of the k_h's, implying that the net sample sizes given by (29) and hence the variance given by (30) are independent of the k_h's. By taking each k_h as large as possible and n_h^{gross} as small as possible, the response rates L_h can be maximized without loss of precision. In the second case the bias is independent of the k_h's. Without bothering about the response rates, one should take each k_h as small as possible and n_h^{gross} as large as possible in order to maximize the net stratum sample sizes and hence to minimize the variance. In general, it is hard to find a reasonable balance between precision and response rate. The easiest way to proceed is to determine the k_h's in advance such that the response rates satisfy some minimum requirements, and to calculate the minimal variance given by (30).

In the following we suggest a slightly more complicated way to find a balance between the total response rate and the minimal variance. To that purpose we need to specify a relation between precision and response fraction. However, since it is nearly impossible to quantify the bias, it is also nearly impossible to quantify the utility of a high response fraction. Therefore, we limit ourselves by treating the response fraction as an ordinal quantity; the higher the response fraction the better. We have shown that both the response rate and the minimal variance depend on the follow up intensities, i.e. $L = L(k_1, \ldots, k_H)$ and $V_{\min} = V_{\min}(k_1, \ldots, k_H)$, and that increasing the follow up intensities will result in a higher response rate and a lower precision. By increasing the follow up intensity in a specific stratum by one, we wish to compare the increase in utility of the response rate with the decrease in utility of the precision. A problem in comparing these changes is that the utilities are measured in different units. This problem can be circumvented by comparing the relative changes. We start with $k_h = 1$ for all h. Calculate

$$V_{\min}^{(0)}(1,\ldots,1) = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\sqrt{a_h(1,\ldots,1)}\right)^2}{C - c_0} - \sum_{h=1}^{H} N_h S_h^2 (1-\rho_h^2) \qquad (32)$$

and

$$L^{(0)}(1,\ldots,1) = \left(\sum_{h=1}^{H} N_h S_h L_h(1)\right) \Big/ \left(\sum_{h=1}^{H} N_h S_h\right). \qquad (33)$$

The variance given by (32) and the total response fraction given by (33) can be considered as a situation with no follow ups. In order to study the effect of one follow up, e.g. in stratum h' we take $k_{h'} = 2$ and $k_h = 1$ for $h \neq h'$, and recalculate the minimal variance and the total response fraction:

$$V_{\min}^{(0)}(1,.,1,2,1,.,1) = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)} \sqrt{a_h(1,.,1,2,1,.,1)} \right)^2}{C - c_0} - \sum_{h=1}^{H} N_h S_h^2 (1-\rho_h^2)$$

and

$$L^{(2)}(1,.,1,2,1,.,1) = \left(N_1 S_1 L_1(2) + \sum_{h=2}^{H} N_h S_h L_h(1) \right) \Big/ \left(\sum_{h=1}^{H} N_h S_h \right).$$

The relative increase in the total response rate and minimum variance are respectively:

$$\frac{\Delta L}{L} = \frac{L^{(1)} - L^{(0)}}{L^{(0)}} \text{ and } \frac{\Delta V_{\min}}{V_{\min}} = \frac{V_{\min}^{(1)} - V_{\min}^{(0)}}{V_{\min}^{(0)}}.$$

Now, we define the effect of one follow up in stratum h' by the relative increase in the total response rate divided by the relative increase in the minimal variance:

$$\frac{\Delta L}{L} \Big/ \frac{\Delta V_{\min}}{V_{\min}} = \frac{\Delta L}{L} \frac{V_{\min}}{\Delta V_{\min}}. \qquad (34)$$

Expression (34) can be calculated for each stratum. The stratum with the largest gain/loss ratio is selected for the follow up if this ratio exceeds some predetermined value larger than zero, say 1. Note that if precision is considered more/less important than response rate this predetermined value should be chosen smaller/ larger than 1 (in Section 14 we give a theoretical justification of this decision rule). For example, suppose that the first stratum has the largest gain/loss ratio and is also larger than 1. The corresponding total response fraction and minimal variance calculated with one follow up in the first stratum can now be regarded as the new starting situation, and it becomes possible to study the effect of a second follow up. As before, this second follow up should be allocated to the stratum with the largest gain/loss ratio provided it is larger than 1. The procedure terminates when none of the gain/loss ratios are larger than 1.

10.3 Exhaustive or nearly exhaustive strata

Special care must be taken if the optimal n_h^{net} given by (29) exceeds $L_h N_h$. Suppose stratum 1 is the only stratum in which oversampling is indicated, i.e.

$$n_1^{net}(k_1, \ldots, k_H) \geq N_1 L_1(k_1)$$

and hence

$$n_1^{gross}(k_1, \ldots, k_H) \geq N_1.$$

In such situations take $n_1^{net}(k_1,\ldots,k_H) = N_1 L_1(k_1)$ and recalculate the optimal sample sizes of the remaining strata:

$$n_h^{net} = (C - c_0 - a_1 n_1^{net}) \frac{N_h S_h \sqrt{(1-\rho_h^2)} \big/ \sqrt{a_h}}{\sum_{h=2}^{H} N_h S_h \sqrt{(1-\rho_h^2)} \sqrt{a_h}}, \quad h = 2,\ldots H.$$

The minimum variance should be calculated with the revised net sample sizes:

$$V_{\min} = \frac{\left(\sum_{h=2}^{H} N_h S_h \sqrt{(1-\rho_h^2)} \sqrt{a_h}\right)^2}{C - c_0 - a_1 N_1} - \sum_{h=2}^{H} N_h S_h^2 (1-\rho_h^2)$$

$$+ \left(\frac{1}{L_1} - 1\right) N_1 S_1^2 (1-\rho_1^2). \qquad (35)$$

We note that the first stratum is exhaustive with respect to the gross sample size. However, due to non response, this stratum can be considered as sampled with respect to the net sample size. To study the effect of an additional follow up in the first stratum we distinguish between the following two situations

$$n_1^{net}(k_1+1,\ldots,k_H) \geq N_1 L_1(k_1+1) \quad \text{and} \quad n_1^{net}(k_1+1,\ldots,k_H) < N_1 L_1(k_1+1)$$

In the first situation one should take $n_1^{net}(k_1+1,\ldots,k_H) = N_1 L_1(k_1+1) \geq N_1 L_1(k_1)$ and recalculate the remaining optimal net sample sizes. We note that in this particular situation the net sample size increases by an additional follow up, contrary to the general situation described in Section 10.1. In the second situation the population size of the stratum is no longer restrictive, and the procedure as described in the previous section can be applied straightforwardly.

10.4 Optimum versus realized allocation

We consider five reasons why the realized net allocation of a stratified sample may not be optimal (in the sense of formula 26) with as a result a loss in precision:

- another allocation than the optimal net allocation is used, e.g. for reasons of convenience,
- the previous samples, on which the estimates for the standard deviations used in the allocation formula are based, are no longer actual due to a dynamic population,
- the standard deviations used in the net allocation formula deviate from the population standard errors, e.g. due to selective non response with respect to the population variances in previous samples,
- unexpected (not according to our model) selective non response (resulting in different response fractions for the strata), in which case L_h deviates from the real response fraction,
- a misspecification of the cost function, e.g. the cost parameters a_h are estimated wrongly.

In order to calculate the optimal net allocation the standard deviations S_{yh} and the response fractions L_h need to be known. However, the optimal allocation is determined before the sample is drawn and hence the standard deviations and response fractions cannot be determined on the basis of the sample. This problem is usually handled by using the results from a similar survey of the previous year, assuming that the variation within the strata and the response fractions will not change much from year to year. When there have been substantial changes within a stratum, for instance due to changes in the population (births, deaths, mergers, etc.), the allocation may no longer be optimal for the sample at hand. Whether the allocation used in a particular survey is (almost) optimal, can be determined by comparing the actual stratum sample size with the 'optimal' stratum sample size as calculated above. The loss in precision resulting from using a less than optimal allocation can be assessed by calculating the difference between the variance for the optimal allocation and the variance for the realized allocation.

11. Trade off between costs and quality

11.1 Overview

To evaluate the relation between accuracy and costs we will study the effect of simulating changes in two instruments that are known to affect precision: the gross sample size and the follow up intensity. In this section we no longer treat costs as a given constant, but as an output variable which may change as the gross sample size and the follow up intensities change. Both an increase in the gross sample size as a higher follow up intensity will result in a greater precision, i.e., reduction of variance. However, the survey costs will also increase. This model will be used in the simulation, which will be discussed in Part C. The results of the simulation will give an indication of the 'marginal costs' of a variance reduction given the process characteristics of the surveys of the Member States. The purpose is to get insight in how differences in the process characteristics between the surveys of the Member States affect the trade off between costs and precision.

As the starting point for our analysis we take the quality indicators with respect to the sampling error described in Section 8 and the cost functions given in Section 9. The main characteristics of the simulation model are summarized below. We distinguish:

<u>Exogenous variables</u>

- the success rates (λ_{hi})
- the use of auxiliary information (as expressed in the correlation coefficient ρ_h^2)

<u>Instruments</u>

- the gross allocation scheme (ξ_h^{gross}) and the gross sample size n^{gross}
- the follow up intensities $k_1, \ldots, k_H$: number of attempts to contact units.

<u>Quality indicators</u>

- the sampling variance $V(\hat{Y}_r)$
- the total response fraction $L(y)$

<u>Cost indicators</u>

- the process costs (C)
- the net response burden (B)

We assume that for the survey process in question information on the model characteristics is available. In the present study only two instruments are considered; the gross sample size and the follow up intensities. Although the field work organization (which determines the values of λ_{hi}) and the use of auxiliary information (which determines the values of ρ_h^2) could also be taken as instruments, here they are assumed to be fixed. With respect to the gross and net allocation scheme we assume that (before changing the instruments) the optimal allocation is used, as described in Section 10.1. Changing the instruments will affect the allocation scheme as well; it may deviate from the original allocation scheme if the two instruments are used independently of each other. Namely, an additional follow up in a particular stratum without changing the gross sample stratum size will affect the net allocation scheme, and the resulting allocation may no longer be optimal (compare equation 31). However, if it is desired that the allocation scheme is also optimal after changing the instruments, then these two instrument should be used simultaneously. In the first case it is possible to study the effects on costs and quality of changing one of the instruments at a time, while keeping the other constant. In the latter case it is more sensible to study the effect on all quality indicators simultaneously by changing the total cost budget and applying the instruments according to an 'optimal' mix (see Section 11.3). To see how the changes affect the quality and cost indicators the relations among the model characteristics need to be specified. This will be the subject of Section 11.2.

11.2 Model specifications

The first step in developing the model is to specify the relationships between the instruments and the quality and cost indicators. Denote $k^t = (k_1, \ldots, k_H)$. From Sections 8 to 10 we summarize the following indicators

$$V(\hat{Y}_r) = \frac{\sum_{h=1}^{H} N_h^2 S_h^2 (1-\rho_h^2)}{L_h n_h^{gross}} - \sum_{h=1}^{H} N_h S_h^2 (1-\rho_h^2)\,,$$

which reduces to

$$V_{\min} = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\sqrt{a_h}\right)^2}{C-c_0} - \sum_{h=1}^{H} N_h S_h^2 (1-\rho_h^2)$$

in case of optimal allocation,

$$L(y) = \left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)} L_h\right) \Bigg/ \left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\right),$$

and

$$C = c_0 + \sum_{h=1}^{H} \left[a_h^3 + a_h^4 (M_h - 1) + (a_h^5 + a_h^6) L_h\right] n_h^{gross}\,,$$

where $L_h = L_h(k_h)$, $M_h = M_h(k_h)$, $n_h^{gross} = n_h^{gross}(k)$, and $a_h = a_h(k)$. We distinguish between two situations, namely 1) the additional follow ups will also affect the gross sample sizes as indicated by equation (31) and 2) the additional follow ups do not affect the gross sample sizes.

11.3 Trade off between costs and quality; situation 1

In this situation we will examine the effect on the quality indicators of increasing the total cost budget by an amount of ΔC. These extra resources can be employed to increase the intensity of the follow up procedure, to increase the gross sample size or in both. The change in minimal variance can be obtained from (30):

$$\Delta V_{\min} = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\sqrt{a_h(k_h+\Delta k_h)}\right)^2}{C-c_0+\Delta C} - \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\sqrt{a_h(k_h)}\right)^2}{C-c_0}. \qquad (36)$$

Since a_h is increasing in k_h, the decrease in minimal variance will be less the more ΔC is employed to increase the follow up intensities. We note that in case of exhaustive (or nearly exhaustive) strata (36) should be adapted, as described in Section 10.3. Let $\Delta L_h = L_h(k_h + \Delta k_h) - L_h(k_h)$ denote the change of the response fraction in stratum *h*, then the increase in the total response rate can easily be obtained from (6)

$$\Delta L(y) = \left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\Delta L_h\right) \Bigg/ \left(\sum_{h=1}^{H} N_h S_h \sqrt{(1-\rho_h^2)}\right). \qquad (37)$$

By means of (36) and (37) we may calculate the 'marginal' profits of an increase in the total costs budget: $\Delta V_{\min}/\Delta C$ and $\Delta L/\Delta C$. We note that these marginal profits depend on the unit of measurement. To avoid this dependence we can calculate the so-called elasticity's

$$\frac{\Delta V_{\min}}{V_{\min}} \Bigg/ \frac{\Delta C}{C} \text{ and } \frac{\Delta L}{L} \Bigg/ \frac{\Delta C}{C}.$$

An increase in the costs budget with an amount of $\Delta C/C \times 100$ percent results in a decrease in the variance with $\Delta V_{\min}/V_{\min} \times 100$ percent and an increase in the response rate with $\Delta L/L \times 100$ percent. As described in Section 10.2 these relative improvements may be the result of an 'optimal' mix between the increase in gross sample sizes and the increase in follow up intensities.

11.4 Trade off between costs and quality; situation 2

In this situation we will study the effect on the quality and costs indicators of changing the gross sample sizes and the follow up intensities independently of each other. To be specific, we no longer strive for an optimal allocation scheme as given by (29). As a result the choice of the gross sample sizes and the follow up intensities are no longer related by (31), i.e. n_h^{gross} is no longer a function of *k*. Instead we fix the gross allocation scheme. As a consequence the resulting net allocation scheme is probably sub-optimal. First we consider a change in gross sample size (holding the follow up intensities constant):

$$\Delta n_h^{net} = L_h(k_h)\xi_h^{gross}\Delta n^{gross}. \tag{38}$$

The change in the variance can easily be obtained from (10):

$$\Delta V(\hat{Y}_r) = \sum_{h=1}^{H}\left(\frac{1}{n_h^{net}} - \frac{1}{n_h^{net} + \Delta n_h^{net}}\right) N_h^2 (1-\rho_h^2) S_{yh}^2 \tag{39}$$

Note that we do not use the minimal variance formula, because the resulting allocation scheme may be sub-optimal. Inserting (38) into (39) gives the resulting change in variance. Again, we note that in case of exhaustive (or nearly exhaustive) strata the changes of the net sample sizes in (38) should be adapted, see Section 9.3. The total response rate does not change, because the response rates L_h only depend on the follow up intensities. The change in the total costs equals

$$\Delta C = \sum_{h=1}^{H}\left[a_h^3 + a_h^4(M_h - 1) + (a_h^5 + a_h^6)L_h\right]\xi_h^{gross}\Delta n^{gross} \tag{40}$$

Next we consider a change in k_h, i.e. a change in the follow up intensity in the *h*-th stratum. The change in the net stratum sample size is (holding the gross stratum sample sizes constant):

$$\Delta n_h^{net} = \Delta L_h \xi_h^{gross} n^{gross} \tag{41}$$

with $\Delta L_h = L_h(k_h + \Delta k_h) - L_h(k_h)$. Inserting (41) into (39) and (37) respectively, gives the change in variance and the change in total response rate. The change in the total costs is

$$\Delta C = \sum_{h=1}^{H}\left[a_h^4 \Delta M_h + (a_h^5 + a_h^6)\Delta L_h\right]\xi_h^{gross} n^{gross},$$

where $\Delta M_h = M_h(k_h + \Delta k_h) - M_h(k_h)$. Eventually, we could calculate the following quotients:

$$\frac{\Delta V(\hat{Y})}{\Delta n^{gross}}, \quad \frac{\Delta L(Y)}{\Delta n^{gross}}, \quad \frac{\Delta C}{\Delta n^{gross}}, \tag{42}$$

and

$$\frac{\Delta V(\hat{Y})}{\Delta k_h}, \quad \frac{\Delta L(Y)}{\Delta k_h}, \quad \frac{\Delta C}{\Delta k_h}. \tag{43}$$

These quotients show the change in the cost and quality indicators due to a change in the instruments. The larger the ratio the more effect an instrument has on the indicator. Again, (42) and (43) can be replaced by their elasticity's in order to eliminate the dependence on the unit of measurement.

11.5 Implementation

In order to implement the model in the computer program, the theoretical relationships between the instruments and the quality and cost indicators will be worked out below. For the implementation some additional assumptions are made as well. For reasons of simplicity no auxiliary information is used, i.e., it is assumed that $\rho_h^2 = 0$. In this case the regression estimator reduces to the direct estimator.

In order to carry out the simulations, reliable information about the statistical process of the Member States is required. Especially, information on the sampling design and the costs of activities is needed.

If no information on the survey costs is available simulations cannot be performed. On the other hand, if all information is available except data characteristics such as the sample mean and the sample variance, one can choose to perform the simulation using the values from similar surveys. For example, the unknown sample variance for survey A can be substituted by the known sample variance obtained in survey B, under the condition that the strata structure is the same. Instead of one survey, one can also use the average of a number of similar surveys as a benchmark for imputation. In the case of substituted data characteristics differences in precision between the surveys merely reflect differences in stratum population and sample sizes.

12. Interrelations between sampling error and frame error

In this section we consider the sampling frame as a list of units, which are classified according to some specific classification variable. The most commonly used classifications in economic statistics refer to activities, institutional sectors, regions, size and products (EUROSTAT, 1998). In Section 8, we discussed the general regression estimator in case of stratified designs, where the stratification variables are derived from one or more classification variables. In addition, due to a continuously changing population, e.g. birth, growth, or death of enterprises, the sampling frame may suffer from undercoverage, misclassification errors, or overcoverage. Undercoverage may have a serious but unknown effect on the bias of the general regression estimator. Without additional research, like inspection of births, it is very hard to quantify this bias. In the following we will concentrate on the effect of misclassification errors and overcoverage on the sampling error.

As long as misclassification errors do not result in undercoverage, the general regression estimator for stratified designs, i.e. formula (9), remains (approximately) unbiased. However, misclassifications may result in higher variances. This can be argued as follows. It is common practice to construct strata so that they are as homogeneous as possible with respect to the target variable *y*, because homogeneous strata will result in low variances (see formula 9 or 30). The strata are constructed by means of the classification variables. In case of misclassification errors the division into homogeneous strata is probably less successful than in case of no misclassification errors. We conclude that classification errors may result in a less effective stratified design and hence in higher variances.

A sampling frame suffering from overcoverage contains units which do not belong to the target population. Such units may turn up in the sample if they cannot be identified beforehand, i.e. before sampling. Now, constructing estimates in case of overcoverage can be considered as constructing estimates for subpopulations. To be specific, both the sampling frame (N units) and the target population (N_a units) are stratified into *H* strata according to the same classification variables. Let N_h denote the stratum total of the *h*-th stratum including the overcoverage and N_{ah} the stratum total of the *h*-th stratum without the overcoverage. Furthermore, denote

$$y'_{hi} = \begin{cases} y_{hi} & \text{if the unit belongs to the target population} \\ 0 & \text{otherwise} \end{cases}$$

and

$$x'_{hi} = \begin{cases} x_{hi} & \text{if the unit belongs to the target population} \\ 0 & \text{otherwise} \end{cases}.$$

The population total of y'_{hi}, i.e. $\sum_{h=1}^{H}\sum_{i=1}^{N_h} y'_{hi} = \sum_{h=1}^{H}\sum_{i=1}^{N_{ah}} y_{hi} = \sum_{h=1}^{H} Y_{ah} = Y_a$ is the parameter of interest. In a stratified simple random sample this population total can be estimated by the following separated regression estimator:

$$\hat{Y}'_r = \sum_{h=1}^{H} N_h \left[\bar{y}'_h + \hat{B}'^{t}_h (\bar{X}'_h - \bar{x}'_h)\right]. \qquad (44)$$

We distinguish between two cases: 1) there is no auxiliary information available and 2) the stratum totals N_{ah} are known. Let n_{ah}^{net} denote the number of observed units belonging to the *h*-th stratum of the target population. In the first case the general separated regression estimator (44) reduces to

$$\hat{Y}'_{(1)} = \sum_{h=1}^{H} N_h \, \bar{y}'_h = \sum_{h=1}^{H} \frac{N_h}{n_h^{net}} n_{ah}^{net} \, \bar{y}_{ah} \equiv \sum_{h=1}^{H} \hat{N}_{ah} \, \bar{y}_{ah}, \qquad (45)$$

where $\bar{y}_{ah} = \frac{1}{n_{ah}^{net}} \sum_{i=1}^{n_{ah}^{net}} y_{hi}$ is the ordinary estimate of the (target) population mean of the *h*-th stratum, based on the net sample size. The variance of (45) can be worked out as (compare Cochran, 1977, form. 2.62)

$$V(\hat{Y}'_{(1)}) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h^{net}} - \frac{1}{N_h}\right)\left(\frac{N_{ah}}{N_h} S_{ah}^2 + \frac{N_{ah}}{N_h}\left(1 - \frac{N_{ah}}{N_h}\right)\bar{Y}_{ah}^2\right) \qquad (46)$$

It is important to note that in case of non response it may be troublesome to estimate N_{ah}, because our estimator $\frac{N_h}{n_h^{net}} n_{ah}^{net}$ may be severely biased upwards if units falling outside the target population do not respond exactly for this reason. In this case the seriousness of the overcoverage will be underestimated, and hence N_{ah} will be overestimated.

In the second case we take $x_{hi} = 1$ for all units; this yields $\hat{B}'^{t}_h = \bar{y}_{ah}$, $\bar{X}'_h = N_{ah}/N_h$, $\bar{x}'_h = n_{ah}/n_h$, and

$$\hat{Y}'_{(2)} = \sum_{h=1}^{H} N_{ah} \, \bar{y}_{ah}. \qquad (47)$$

If we condition on n_{ah}^{gross}, then the sampling design can be considered as stratified simple random sampling with gross stratum sample size n_{ah}^{gross}. This results in the following formula for the variance

$$V(\hat{Y}'_{(2)}) = \sum_{h=1}^{H} N_{ah}^2 (\frac{1}{n_{ah}^{net}} - \frac{1}{N_{ah}}) S_{ah}^2 \ .$$

For large sample sizes we have approximately (assuming no bias) $n_{ah}^{net} \big/ n_h^{net} \approx N_{ah} \big/ N_h$ and

$$V(\hat{Y}'_{(2)}) \approx \sum_{h=1}^{H} N_h^2 (\frac{1}{n_h^{net}} - \frac{1}{N_h}) \left(\frac{N_{ah}}{N_h} S_{ah}^2 \right). \qquad (48)$$

A comparison between (46) and (48) shows that the reduction of variance due to a knowledge of N_{ah} is larger when the proportions of overcoverage $(1 - N_{ah} \big/ N_h)$ are large.

If we compare the variances (46) and (48) with the variance which would have been obtained in case of no overcoverage

$$V(\hat{Y}_{HT}) = \sum_{h=1}^{H} N_{ah}^2 (\frac{1}{n_h^{net}} - \frac{1}{N_{ah}}) S_{ah}^2 = \sum_{h=1}^{H} N_h^2 (\frac{N_{ah}}{N_h} \frac{1}{n_h^{net}} - \frac{1}{N_h}) \left(\frac{N_{ah}}{N_h} S_{ah}^2 \right),$$

we see that the increase in variance due to overcoverage can be explained by two effects. The first effect is a reduction of effective net sample size; sampled units which do not belong to the target population should be neglected. The second effect is a probable loss of information about the stratum population totals N_{ah}.

If we want to limit the disadvantageous effects of overcoverage we need to know N_{ah}, or even better, we need to identify the misplaced units. When this information is lacking, it may be cheaper to start with a relatively large preliminary sample in which only x'_{hi} is measured. The purpose of this sample is to provide preliminary information with respect to overcoverage. However, this means that the size of the sample in the main survey on y_{hi} must be decreased. This technique is known as double sampling or two-phase sampling (see Cochran, 1977, Chap. 12). One may distinguish between double sampling for proper stratification and double sampling for regression estimates. The use of a preliminary sample can be incorporated into our model. We will not elaborate on this topic.

13. Incorporating net response burden

Until now, the net response burden has been neglected as a cost indicator. Besides costs made by the statistical agency there are costs made by businesses in completing the questionnaires of the business surveys. The total costs made by the responding businesses, the net response burden, is measured as the time spent to fill in the questionnaire. That is, we define the net response burden as

$$B = \sum_{h=1}^{H} \bar{t}_h^{resp} n_h^{net} \ ,$$

where $\bar{t}_h^{resp}$ is the average time to fill in the questionnaire. Note that according to this definition, we are able to discriminate between strata with respect to the length of the questionnaire, taking into account the structure of the sample vis-à-vis the size of the enterprise. In case of fixed costs and optimum allocation, the net response burden can be considered as a function of the follow up intensities $B = B(k)$. We have argued that n_h^{net} is non-increasing in k_h if $\lambda_{h1} \geq \lambda_{h2} \geq \ldots \geq \lambda_{hk_h}$ (see Section 9.1), implying that the net response burden is also non-increasing in k_h (with $k_1, .., k_{h-1}, k_{h+1}, .., k_H$ fixed). One way to incorporate the net response burden into our model is to consider it as a quality indicator instead of a cost function, and to extend the theory of Section 9.2. This will be the subject of the next section.

14. Utility functions

In order to incorporate the net response burden we have to specify the utility function, which is implicitly used in Section 10.2 to relate the minimal variance and the total response rate. Denote this function by $U = U(L,V)$. Since both *L* and *V* are functions of *k*, *U* is also a function of *k*. It is important to note that it is impossible to measure *U*, because *L* and *V* are in general completely different quantities with completely different units of measurement. Instead, we treat *U* as an ordinal variable. For example consider the following utility function

$$U = \log(L) - \log(V). \tag{49}$$

The idea is to maximize *U* as a function of *k*. If the k_h were continuous, we could find this maximum by differentiating *U* with respect to the k_h and setting the partial derivatives to zero:

$$\frac{dU}{dk_h} = \frac{1}{L}\frac{dL}{dk_h} - \frac{1}{V}\frac{dV}{dk_h} = 0, \quad h = 1, \ldots, H .$$

This gives $\frac{1}{L}\frac{dL}{dk_h} = \frac{1}{V}\frac{dV}{dk_h}$ for $h = 1, \ldots, H$. However, in our model the k_h's are discrete. In Section 10.2. we have in fact given an algorithm to approximate the optimal value of *U*. Now, the utility function given by (49) opens the door to incorporate the net response burden. Consider the following extension of this function:

$$U = w_L \log(L) - w_V \log(V) - w_B \log(B) .$$

The weights *w*, indicating the relative importance of each quality indicator, should be determined by the survey manager,. For example, if only precision is considered important (i.e. the non response is considered to be non-selective and the net response burden can be neglected), then one should take $w_V = 1$ and $w_L = w_B = 0$. We note that variance and net response burden are in fact disutility's . Therefore these components have a negative sign. Based on this extended utility function one may search for an optimal follow up strategy, e.g. as described in Section 10.2.

Part C. Manual for the computer program

Introduction

Based on the model developed to assess the quality of structural business statistics of the EU Member States a computer tool is developed. This tool handles the input of the data obtained from the questionnaire (see appendix D) and produces the output as mentioned in Part B. To this part is referred for the theoretical background of the model. In this manual, the use of the data quality program will be explained. The options and possibilities of the program will be discussed and some examples will be given. A case (Utopia) is worked out (see the answers in Appendix D) to illustrate the input and output of the program.

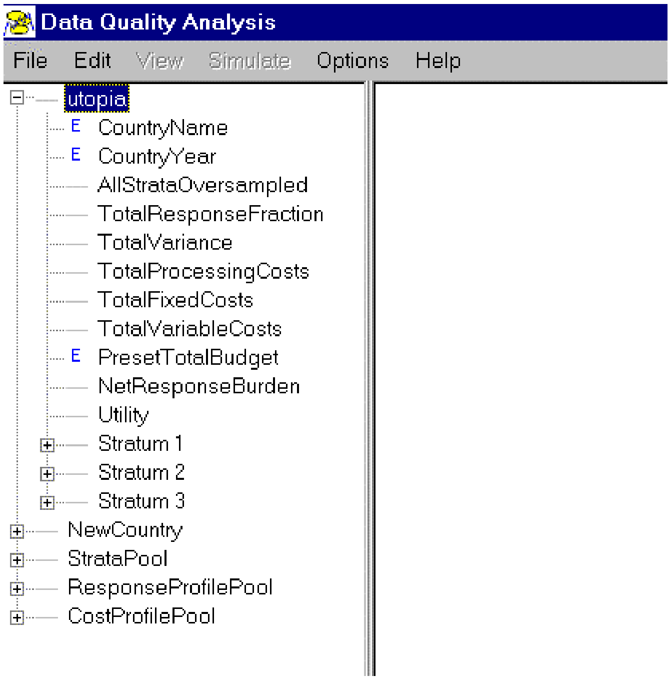
The program is written in the language Delphi 3.02 and runs on a Pentium 16Mb computer. All the input and output is handled by Delphi: the input tables have been made in Paradox and can be examined using the Database Desktop program. The screen is divided into 2 parts: the left part contains a description of the data; the right part is used to display either extra information on the data or to display the result of numerical simulations. This result can be displayed either as text or as a graph. The manual consists of 3 parts: the first part discusses the way the data has been represented in the program, the second part lists all possibilities to edit the data and the third part discusses all menu items.

Note: before starting up the program, the database location must be defined in the *Quality.ini* file (in the same directory as the program), i.e. altering the text after `DataBaseDir = .` This can be done using a simple text editor.

The data structure

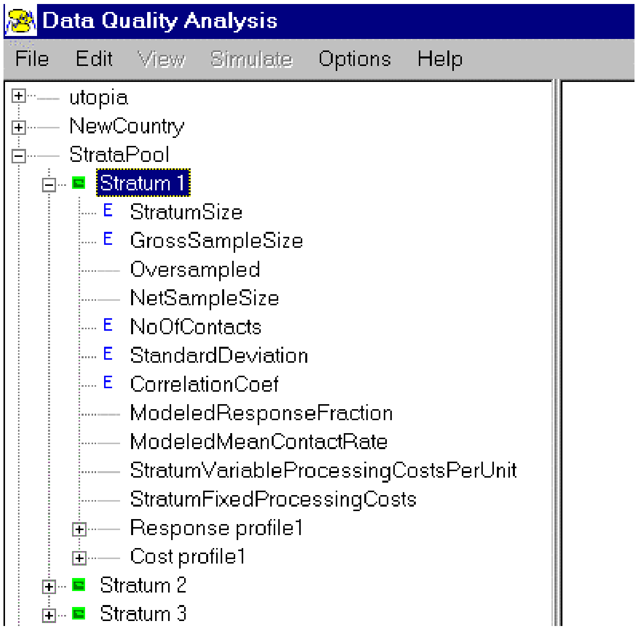
Country:

The data consists of countries (Utopia, etc.) with a number of properties, such as TotalVariance, etc. Attached to the countries are a number of strata (for Utopia these are stratum 1, 2 and 3). The E in front of CountryName and CountryYear indicate that these items are editable by clicking the right mouse button. However, most of the properties of the country are not editable, but will be calculated through the menu items under *Simulate*.



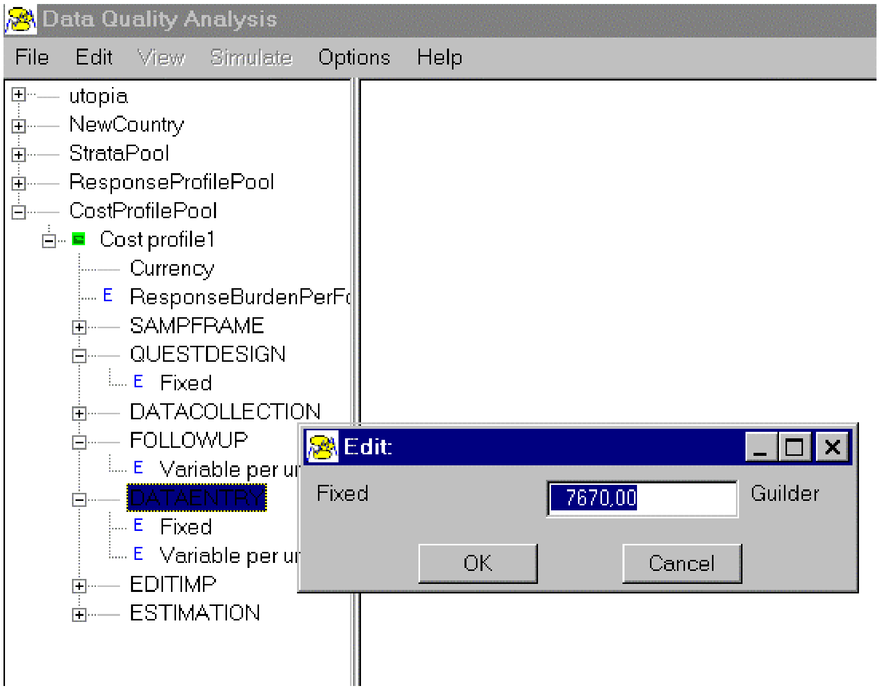
Stratum:

Each stratum consists of a number of properties such as GrossSampleSize, NetSampleSize, etc. and includes 2 profiles: a cost profile and a response profile.



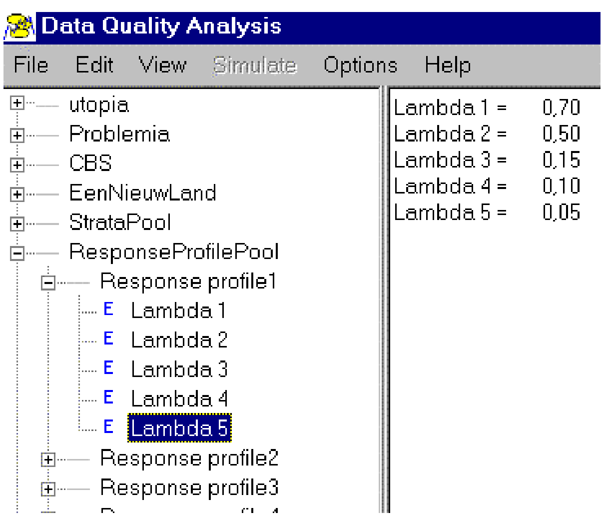
Cost Profile (see Section 9):

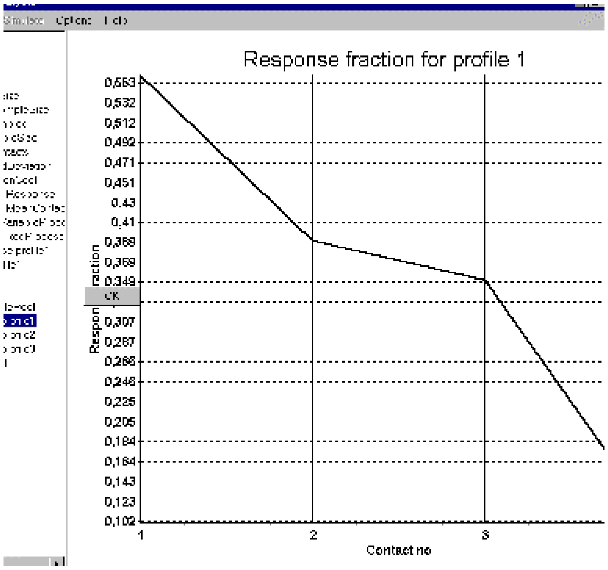
The Cost profile contains 7 cost types; for each of these types 1 or 2 quantities are available: fixed and/or variable costs. If the right button is clicked on editable item (marked by an E) then a small edit window appears.



Response Profile (see Section 8):

The Response Profile contains the response fractions as defined in Section 8. Left clicking on the lambda labels will print their values in the right window. Dragging a response file the right window will show a graph of the response fractions (the lambda's) as a function of the number of contacts.





The pools:

The strata, the response profiles and the cost profiles are organized into pools (which in fact correspond with the tables *variance, respfrac* and *costs*, respectively). ONLY the elements of the pools are editable! This means that in order to edit stratum 1 from country Utopia one has to edit stratum 1 from the StrataPool. Note that this edit will also affect other countries that refer to Stratum 1. In order to prevent this from happening one should define separate strata for both countries in the stratapool.

Editing the data

Before discussing all editing possibilities, it is important to notice the following: all edits are made to the data in the computer. Nothing is definite until stored by using *DB* --> *Tree* or by exiting the program and clicking yes to the question whether to save the data or not.

There are two exceptions to this:

1. If a form is used to edit a country or a stratum then the program data for that country or stratum is saved to the database; once the editing of the form has been completed, the database record (-not all- of the database) is copied back to the program data.
2. If any of the items under 'View' is used, the data being displayed in the form will be saved from the program to the database before displaying the form.

To add an item:

There are 5 items in the file menu:

New Country

New Stratum

New ResponseProfile

New ResponseProfile Entry

New CostProfile

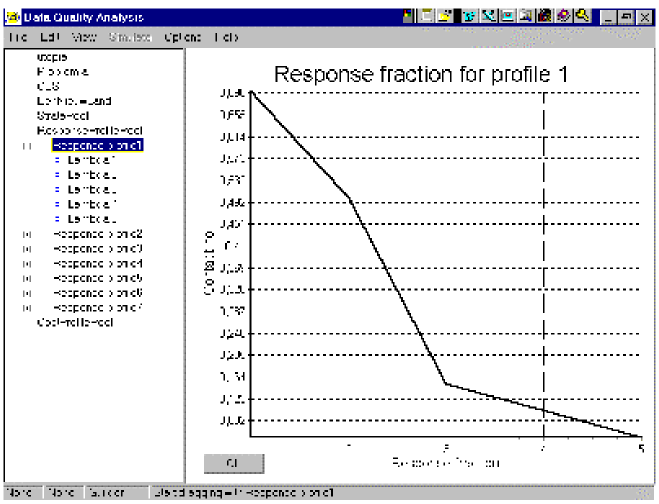
Before selecting this item, one can select corresponding country, etc. from which to copy the data. If none is selected the first country, stratum, etc. from the tree or pool is selected. If the pool does not yet contain any items, a completely new one is generated. The item 'New ResponseProfile Entry' adds a new response fraction (lambda) to the selected response profile.

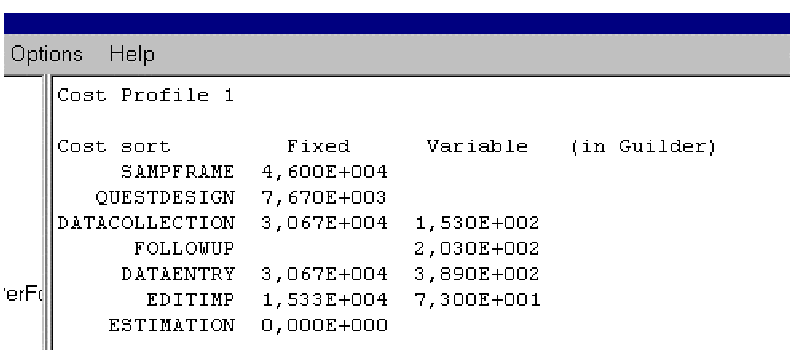
To delete an item:

Select the item and choose *Edit - Delete*. An item can only be deleted if there are no dependent parents; for example, if stratum 1 contains cost profile 1, then cost profile 1 can**not** be deleted from the pool.

Show property value:

Simply click on the relevant item. This will print the property name and value in the right window. For profiles, strata and countries it is possible to drag them to the right area and drop them there; this will cause (part of) their information to be printed there or a graph depicting the information. Below are shown the response and the cost profile, respectively.





Edit property value:

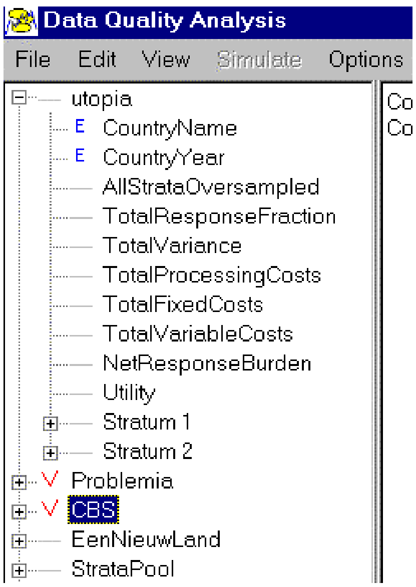
Right click on an item; an item is editable if marked by an E. In case of cost distributions some extra actions are performed: if the value of a FixedFraction or total costs changes, new values for fixed and variable costs are recalculated.

Edit values using a form:

Right click on an item; whether an item is editable by a form is marked by an little form-like icon.

(De)selecting a country for analysis:

Right click on the country; a red check mark will (dis)appear. Notice that the menu item View is only applicable when at least 1 country is selected; the menu item Simulation is only applicable when exactly 1 country is selected.



Adding a new stratum to a country:

Drag the stratum from the stratapool to the country. Note that a tree item is 'dragable' if the mouse icon changes to a hand shape when maneuvering over the item. When dragging, mouse shape changes to an arrow with a small rectangle when crossing allowed target items (in this case the countries). Releasing the mouse button on such a target item will then add the stratum to the country, in this case. Note that it is possible to drag the same stratum more than once to a given country; this will result in a number of identical strata for that country.

Adding a new response profile to a stratum (in the pool):

Drag the response profile to the stratum in the stratapool.

Adding a new cost profile to a stratum (in the pool):

Drag the cost profile to the stratum in the stratapool.

Menu items

File - DB --> Tree:

Copy the contents from the database to the program data.

File - Tree --> DB:

Copy the program data to the database. Using *File – Exit* when exiting the program, the program will ask if this action should be performed or not.

File - New Country, File - New Stratum, etc:

See above: *To add an item*.

File - Copy:

Copy the left window (text or graph) to the clipboard; other window programs can now use the text or the graph using Paste.

File - Export:

Save the contents of the left window.

File - Print:

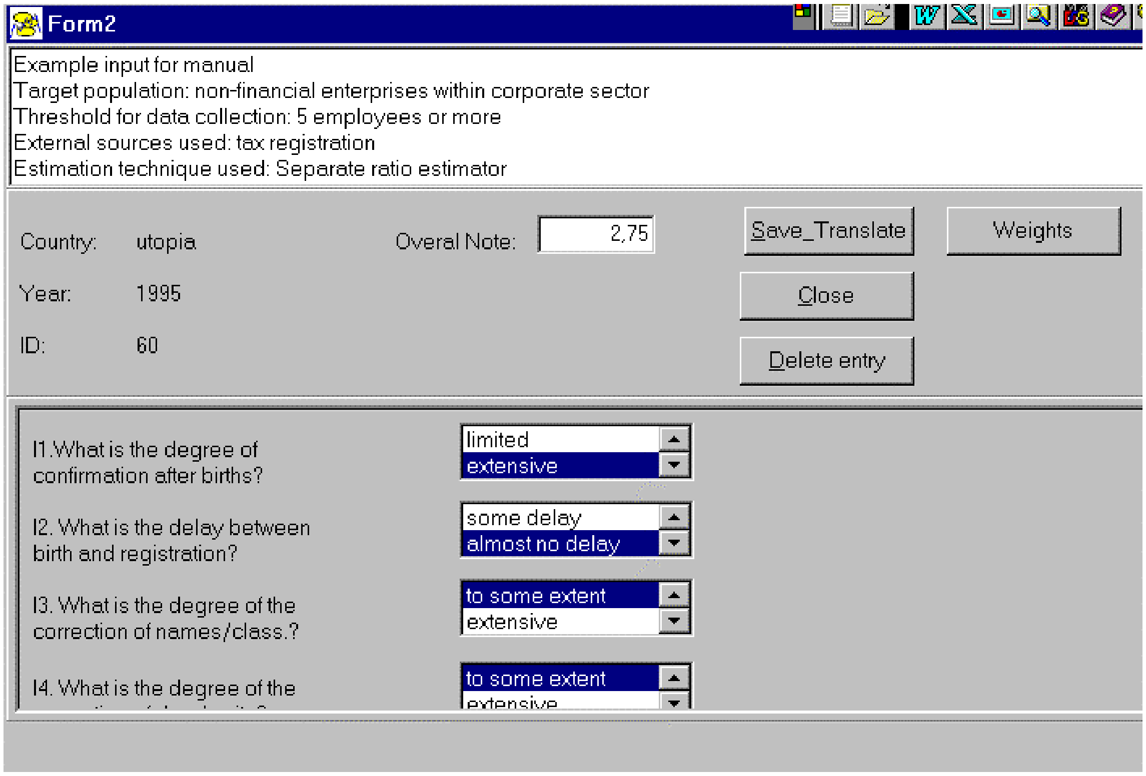
Print the contents of the left window.

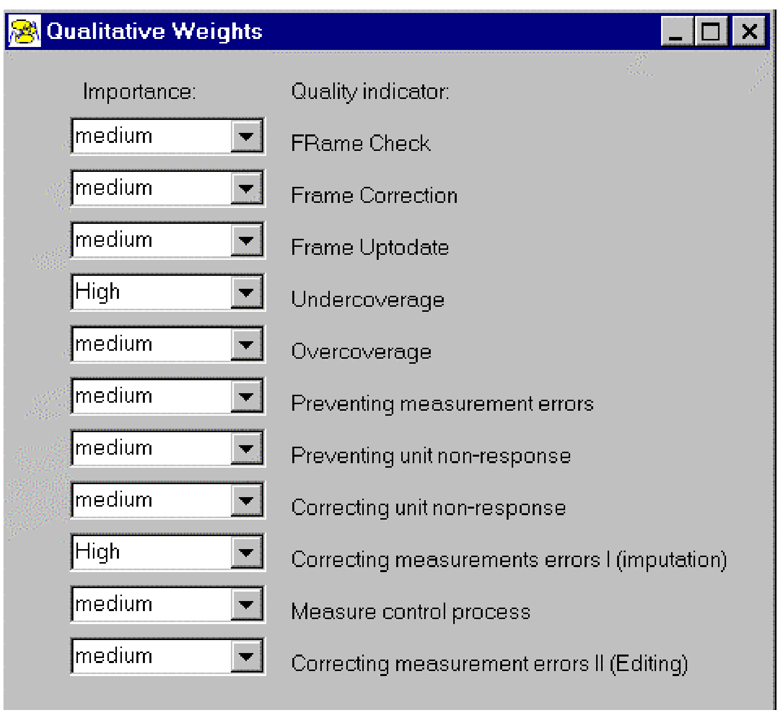
File - Printer setup:

Define the printer and its characteristics.

Edit - Qualitative (Section 7):

Edit the qualitative data obtained from the questionnaire for a selected country. When all the questions of the input list are answered, pushing the button Translate will translate all answers into scores on the quality indicators of the non-sampling errors. All the **Not known/Blank** scores will be added up and will give you a score on the quality indicator 'process control'. If not all questions are answered one can only close the input list. The answers will be saved and the next time the file is opened the unanswered questions can be completed. In addition, it is possible (by pressing the button weights) to edit the relative importance for each of the quality indicators.





Edit - Delete:

Delete the selected tree item; see above: *To delete an item.*

Edit - Clear edit window:

If the left window contains text, its contents can be cleared using this menu item.

View - Performance:

To compare the scores of several countries on the quality indicators for the non-sampling errors, one can select this menu item. We distinguish three types of non-sampling errors: frame errors, non-response errors and measurement errors. Each country receives a score on how well it prevents, checks and corrects for each type of error. This score is indicated by a number of stars. One star will indicate poor quality, two stars will indicate moderate quality, three stars will indicate good quality and four stars will indicate high quality. Using the weights defined previously, these 3 categories will receive a note between 1 and 4. Finally, the overall note based on the same weights is calculated.

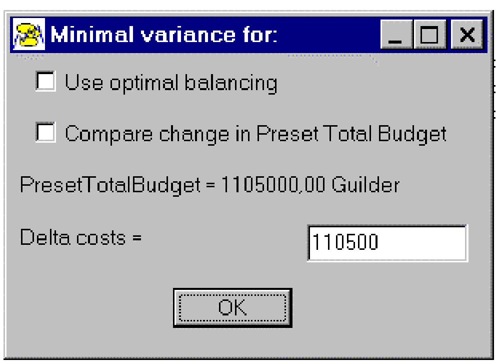
Simulate - Current (Section 8):

Calculate a number of variables at the stratum and country level for the given country 'configuration'. Text output on a number of variables is displayed. In addition a so-called 'result country' is added to the list of countries, which allows to further query any other calculated value. Note that these 'result country' are not editable and will not be save to the database.

Simulate - Min Var (Section 10.1):

Calculate a number of variables at the stratum and country level for the given country 'configuration' using Neyman allocation at the given fixed total budget. There are 2 extra options:

1. In addition to optimal allocation, optimal balancing can be used as well (see Section 10.2).
2. It is also possible to increase the fixed budget by a certain amount and compare the situations with the 2 budgets, i.e. the result of the budget change on n_{net} and k_i .



Simulate - dV/dC and dL/dC (Section 11.3):

(using Neyman allocation at the given fixed total costs)

Obtain the quantities (dV/V) / (dC/C) and (dL/L) / (dC/C). This is done by increasing the number of contacts for each of the strata by 1 and calculating dC, dV and dL as a result of that. These quantities give information on how effective an increase in budget would be spent on V or L.

Simulate - d../dk and d../dn_gross (Section 11.4):

(NOT using Neyman allocation at the given fixed total costs)

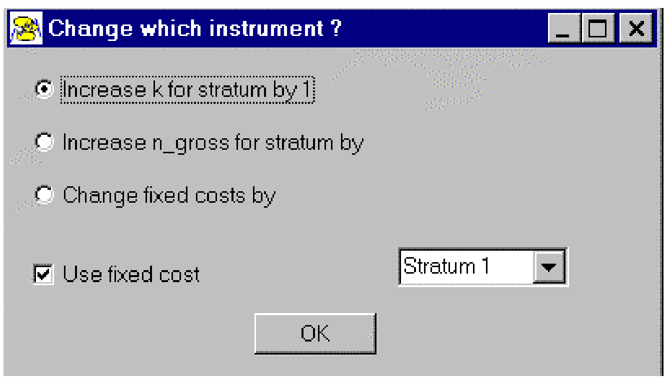
The effect of an increase of the number of contacts (k_h = NoOfContacts) or the gross sample size (n_{gross} = GrossSampleSize) on the following quantities V (TotalVariance), L (TotalResponseFraction) and C (TotalProcessingCosts) is calculated.

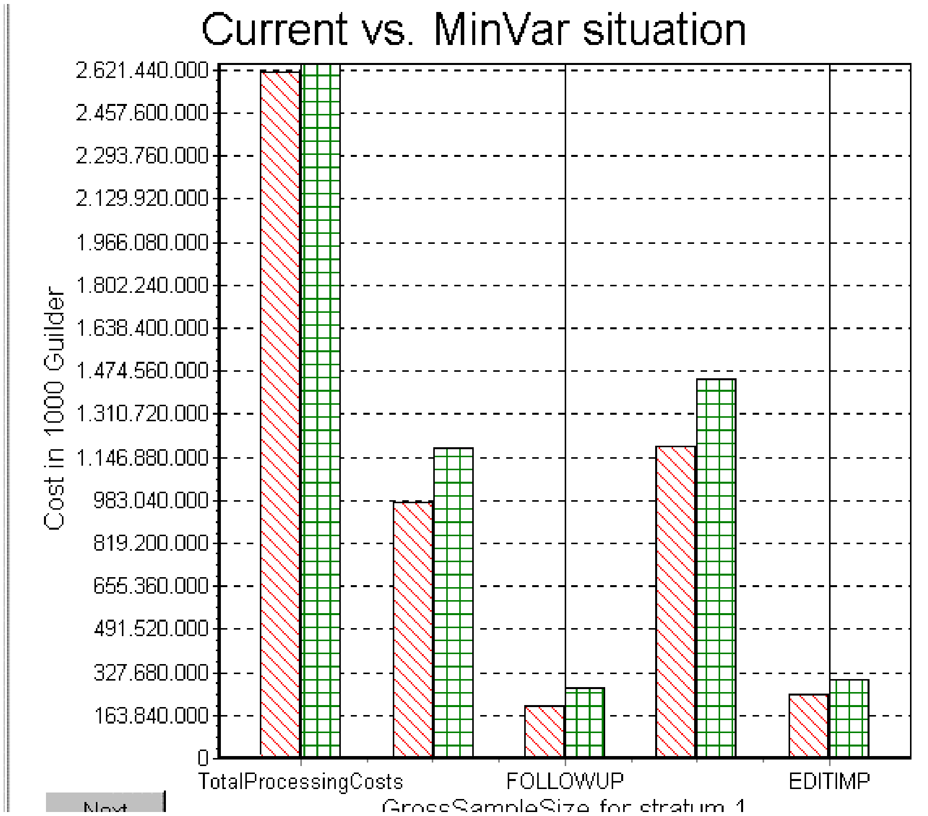
Simulate - Utility (Section 14):

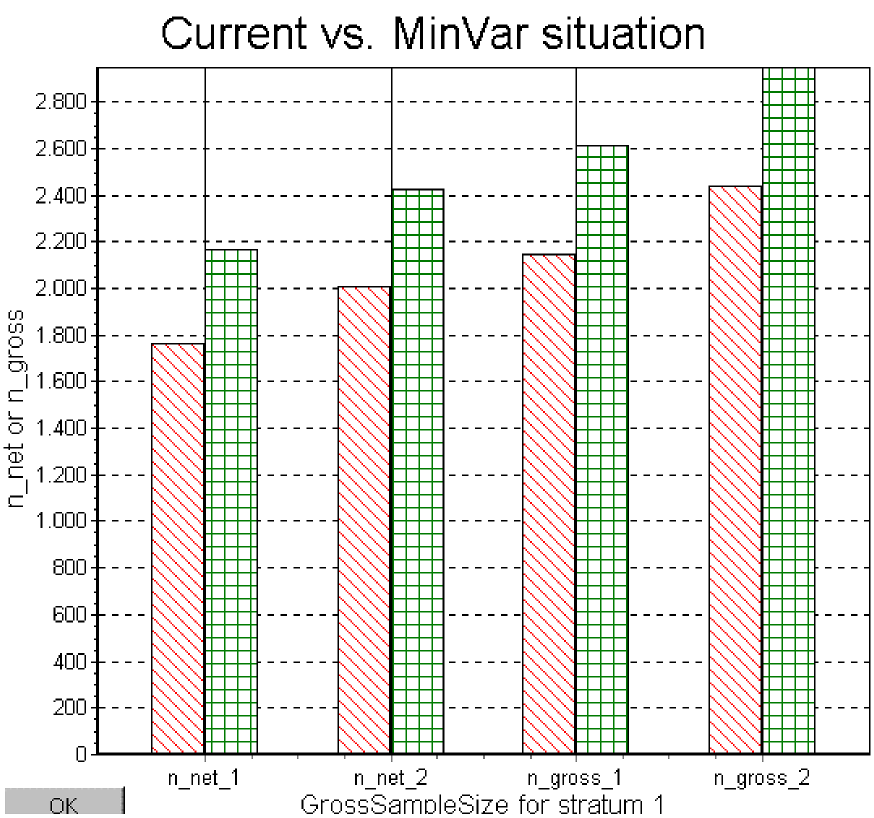
Calculate the Utility function as defined in section 14. The weights (which are default 1,0) can be altered through the options menu item.

Simulate - Compare current and min-var...:

For the selected country the situation before and after a change in k_h, n_{gross} or C can be depicted. After selection of this menu item, a popup window appears which allows the selection of the change and the size of the change. In addition, for changes in k_h,and C one can select with or without fixed cost (= total budget). After OK has been pressed, two graphs are shown: the first shows the changes in costs, the second one shows the changes in n_{net} and n_{gross}; for both graphs the first bar shows (red) shows the initial value and the second bar (green) shows the new value.

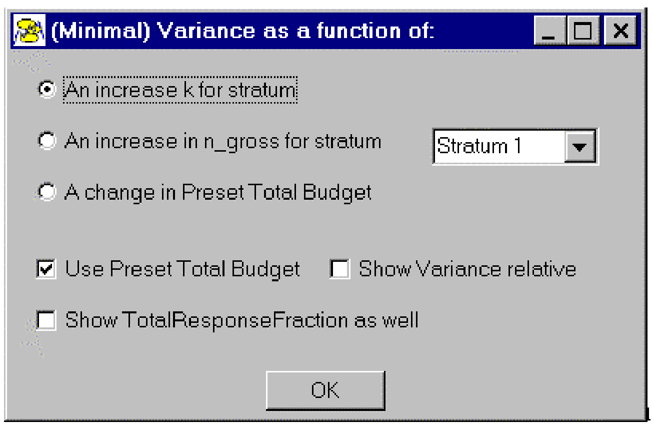


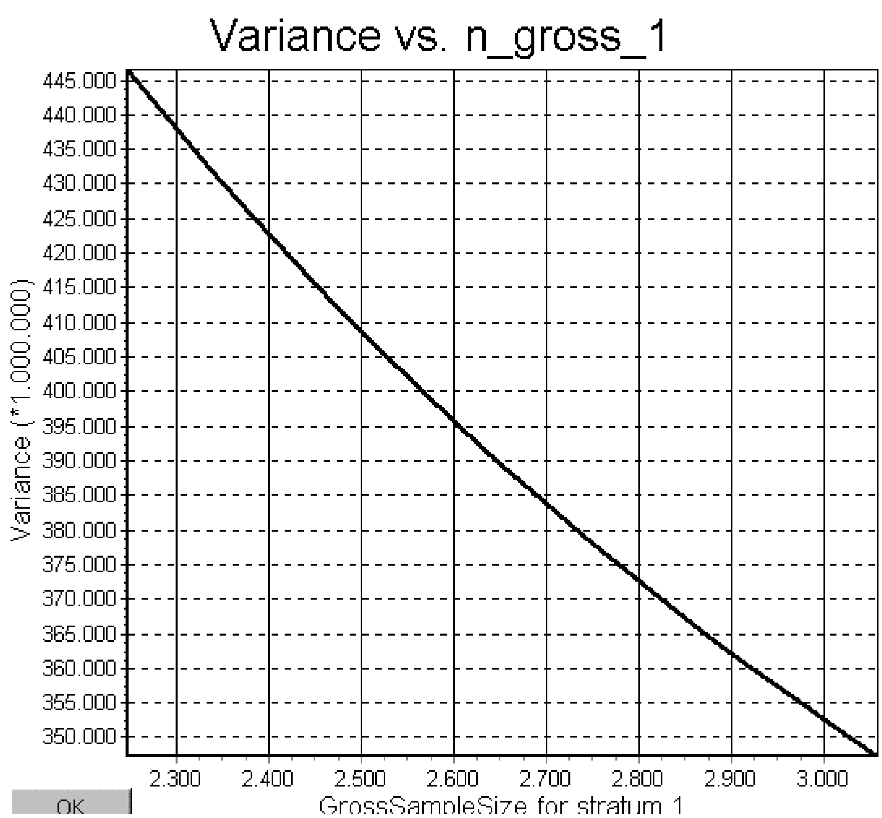

Current vs. MinVar situation
Cost in 1000 Guilder
2.621.440.000
2.457.600.000
2.293.760.000
2.129.920.000
1.966.080.000
1.802.240.000
1.638.400.000
1.474.560.000
1.310.720.000
1.146.880.000
983.040.000
819.200.000
655.360.000
491.520.000
327.680.000
163.840.000
0
TotalProcessingCosts
FOLLOWUP
EDITIMP


Current vs. MinVar situation
n_net or n_gross
2.800
2.600
2.400
2.200
2.000
1.800
1.600
1.400
1.200
1.000
800
600
400
200
0
n_net_1
n_net_2
n_gross_1
n_gross_2
OK
GrossSampleSize for stratum 1

Simulate - (Min)Var as a function of...:

For the selected country (minimal) variance as a function of k_h, n_{gross} or C can be depicted. After selection of this menu item, a popup window appears which allows the selection of the change and the size of the change. In addition, for changes in k_h,and C one can select with or without fixed cost (= total budget). After OK has been pressed, a graph is shown of (minimal) variance versus the selected variable.





Options:

Utility weights: These correspond to the weights as defined in Section 14.

Simulation parameters: Currently there is only one parameter that can be set: the threshold value for the ratio (dL/L)/(dV/V) (see Simulate - (dL/L)/(dV/V) (Section 10.2)):

Currencies: the current currency which is used for display and conversions rates between currencies.

Conclusions

In accordance with the project objectives (Statistics Netherlands, 1995) we have developed a model relating accuracy to the survey costs and administrative burden. We have focused primarily on the precision of the estimates as a measure of quality associated with sampling errors. Besides the variance we have incorporated the response fraction as a quality measure for the bias due to non response. The model makes it possible to carry out a simple sensitivity analysis to examine the trade off between accuracy and costs under different assumptions about optimality of the allocation scheme.

In addition, we considered non sampling errors by defining relatively simple quality indicators that can easily be applied in practice. An example has been worked out to illustrate the transformation of the process information (obtained in a questionnaire sent to the statistical agencies in the Member States) into values on a number of descriptive quality indicators. A comparison of these scores for different surveys having different process characteristics and using different methods will give us some of idea of the possibilities for quality improvement.

In Section 10 the optimal allocation is calculated as the most effective survey strategy given the total costs of a particular survey. In Section 11 we have modelled the trade off between costs and quality taking into account the survey methods as they are used in practice. By manipulating the sample size at two points in the survey process (in the sampling design and in the follow up procedure) the effects on the precision and the survey costs are examined. Comparing the results for the surveys of the Member States gives an impression of the effect of alternative methods on the quality-cost relation.

The model is implemented in a software tool. For the model to work, many model parameters must be given values that coincide as much as possible with the situation encountered in practice. As the input of the model is rather open, the user of the program can restructure the information on data and process characteristics to some extent to close the gap between theory (model) and practice.

In conclusion, we hope that our model will contribute to the discussion about quality improvement of structural business statistics in the Member States. Referring to the official project objective (see Statistics Netherlands, 1995), the model may be regarded as a starting point for determining an optimal and cost effective practice to obtain accurate estimates.

Appendix A. Proof that *M*/*L* is non-decreasing as a function of *k*

We have to show that that $M_h(k_h)/L_h(k_h) \geq M_h(k_h-1)/L_h(k_h-1)$ if $\lambda_{h1} \geq \lambda_{h2} \geq \ldots \geq \lambda_{hk_h}$. To that purpose we define $b_0 = 1$ and $b_j = \prod_{i=1}^{j}(1-\lambda_i)$ for $j = 1, \ldots, k-1$ and note that $M(k) = \sum_{j=0}^{k-1} b_j$ and $L(k) = \sum_{j=0}^{k-1} \lambda_{j+1} b_j$ (for notational convenience we have omitted the index *k*). Now, a comparison between

$$\frac{M(k)}{M(k-1)} = \frac{\sum_{j=0}^{k-1} b_j}{\sum_{j=0}^{k-2} b_j} = 1 + \frac{b_{k-1}}{\sum_{j=0}^{k-2} b_j} \text{ and}$$

$$\frac{L(k)}{L(k-1)} = \frac{\sum_{j=0}^{k-1} \lambda_{j+1} b_j}{\sum_{j=0}^{k-2} \lambda_{j+1} b_j} = 1 + \frac{\lambda_k b_{k-1}}{\sum_{j=0}^{k-2} \lambda_{j+1} b_j} = 1 + \frac{b_{k-1}}{\sum_{j=0}^{k-2} \frac{\lambda_{j+1}}{\lambda_k} b_j}$$

yields that $\frac{M(k)}{M(k-1)} \geq \frac{L(k)}{L(k-1)}$ since $\frac{\lambda_{j+1}}{\lambda_k} \geq 1$ for all $j = 0, \ldots, k-2$ and $b_j \geq 0$ for all $j = 0, \ldots, k-1$. In fact, by induction we have shown that the assertion holds for all $k \geq 2$.

Appendix B. Advance letter

Heerlen, 13 february 1997

Dear colleague,

Referring to the new regulation on structural business statistics, Member States are expected to provide Eurostat with information on the quality of their survey results. A proposal for a quality report on business statistics is currently under discussion at the EEA Working Party Meetings on 'Assessment of the quality of business statistics'. In order to get comparable and reliable outcomes for business surveys from different statistical agencies, means have to be found for controlling the accuracy of the estimates. For this reason Eurostat has initiated the project 'Data Quality', which is carried out by Statistics Netherlands within the framework of SUPCOM.

The objective of the project is to develop a model by which the trade-off between the accuracy of the survey results and the survey costs can be evaluated given the specific properties of the business surveys in the different Member States. We expect that the model can give us more insight in the differences in the survey process characteristics between the Member States and how these differences affect the survey results. We need the help of the Member States to get sufficient information about the characteristics of their survey process. We are interested in two types of information:

(a) data characteristics, in order to be able to calculate quantitative indicators like the standard errors

(b) process and cost characteristics for evaluating the survey results in quantitative and qualitative terms.

In this research project we will limit ourselves to two structural business statistics:

(a) Manufacturing Industry (section D of the NACE Rev.1), and

(b) Wholesale and Retail Trade, Repair of motor vehicles and motor cycles and personal and household goods (section G of the NACE Rev.1).

Furthermore, the quantitative quality measures will be restricted to some key variables of the surveys, as specified in the appendix. To obtain information on the process characteristics we have developed a questionnaire. On the whole this questionnaire has a similar scope regarding the assessment of accuracy as is agreed upon in the Working Group on quality.

We would appreciate it very much if you could help us with our project. We realise that it will take you some time to collect the requested information. However, we hope the results of our project will also be of interest to you. For instance, we expect this project will be helpful to clarify some of the problems faced in practice when assessing the quality of business statistics. Furthermore, it is worth knowing what differences in process characteristics exist between the Member States, and how they affect quality. It goes without saying that the more information we can get from the Member States, the more our model can contribute to the discussion on quality at Community level.

To be of use for us, we need to have the information by April 15, 1997. We are more than willing to help you answer the questionnaire. Please let us know if there are any problems. For all your questions about the project and the questionnaire you can contact the project leader of the Data Quality project.

Yours sincerely,

René Huigen

Statistics Netherlands

Appendix C. Data characteristics

In the questionnaire you are asked to specify the sample design of the survey in question. We implicitly assume that all business surveys are drawn according to a stratified sample. To be able to quantify the sampling error, we also have to know some descriptive measures of an actual data set of the survey. For each stratum we have to know the following descriptive measures:

(1) the net sample size,

(2) the sample totals of the key variables listed below

(3) the sample variances of the key variables listed below i.e., $S_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(y_{h_i} - \bar{y}_h \right)^2$

If auxiliary information is used, we are also interested in the sample (multiple) correlation coefficients between the key variables and the auxiliary variables.

For our study, the data characteristics influencing quality and cost are of primary interest, and not so much the exact values of the variables for a given year. So, it is not necessary to obtain these descriptive measures for exactly the same data set as the one prepared for Eurostat. What is important is that the measures are typical for the data set of the survey in question.

Referring to the appendix of the new regulation on structural business statistics, the quantitative indicators of quality will be restricted to the following key variables for both Manufacturing and Domestic Trade and Repair:

- turnover, total
- value of goods produced
- value added incl. VAT
- value added exc. VAT
- operating surplus
- purchases, total
- costs of goods purchased for resale as purchased
- labour costs
- number of employed persons
- number of employees

Appendix D. Questionnaire

Remark: the coding to the pre-defined answer categories is not stated on the original questionnaire.

The answers refer to a fictive case (Utopia) serving as example for the Data Quality computer program.

Questions for statistical agencies about their business surveys

Name of the annual survey: **UTOPIA**

Year of the survey: **1995**

Population

Q1: Please give a complete description of the target population, i.e. the conceptual population of units at which the survey is aimed? *Non-financial enterprises within the corporate sector*

Q2: What is the threshold for data collection with respect to the size of businesses covered by the survey? *Enterprises with 5 or more employees*

Q3: In case of a threshold for data collection, have any external sources been used to obtain information about the size classes not included? *Tax registration*

Q4: Please indicate and describe if necessary which restrictions or exclusions (other than those mentioned for Q2) apply to the target population? *None*

Q5: Please describe the register(s) used for the industrial/business surveys? *Business register*

Business register/Sampling frame

Some countries are using a central business register as sampling frame, other countries are using several registers. The questions Q6 - Q13 apply to the main register(s) used for the industrial/business survey.

Maintenance of the main register

Identifying births

Q6: Is any confirmation procedure carried out before potential births of population units are recorded on the register?

[ ] No

[x] Yes. Please, describe: *All new enterprises are first checked with the chambers of commerce and with the tax registration*

coding: none, limited, extensive, (not known/blank) [0, 1 ,2 ,0]

Q7: How long on average is the delay between birth and inclusion of a population unit in the business register?

| | 5 | weeks

[ ] not known

coding: some delay , almost no delay, (not known/blank) [1 ,2 ,0]

Updating contact, classification, structure and size information of businesses

Q8: Please indicate what kind of sources are used to update names, addresses, size indicators, classification codes and the structure of businesses in the register? If possible, specify or estimate the percentage of updates accounted for by each source (multiple answers possible).

	Percentage
[x] Feedback from surveys using the register	*90%*
[ ] Quality surveys specially conducted to maintain the register	
[x] Administrative sources	*10%*
[ ] Account management (profiling of businesses by personal contact)	
[ ] Other, namely:	

coding: to some extent, extensive, (not known/blank) [1, 2, 0]

Identifying deaths

Q9: Please indicate what kind of sources are used to identify 'dead' population units in the business register? If possible, specify or estimate the percentage of deaths identified by each source (multiple answers possible).

	Percentage
[x] Feedback from surveys using the register.	*13%*
[ ] Quality surveys specially conducted to maintain the register.	
[x] Follow up of nonrespondents.	*2%*
[x] Administrative sources.	*85%*
[ ] Account managment	
[ ] Media	
[ ] Other:..	

coding: to some extent, extensive, (not known/blank) [1, 2, 0]

Identifying duplicates

Q10: Please indicate how duplicate units in the business register are identified? If possible, specify or estimate the percentage duplicates identified by each source (multiple answers possible).

	Percentage
[ ] Feedback from surveys using the register	
[ ] Quality surveys specially conducted to maintain the register	
[ ] Sophisticated matching procedures	
[ ] Other, please describe..	

coding: to some extent, extensive, (not known/blank) [1, 2, 0]

Q11: If special quality surveys are carried out to maintain and update the register, please specify

a) the sampling fraction ....

coding: none, small sample, large sample, (not known/blank) [0, 1, 2, 0]

b) the frequency with which these samples are taken ....

coding: never, once a year, more often, (not known/blank) [0, 1, 2, 0]

Sampling frame

Coverage

Q12: No register used as sampling frame can claim to provide a complete, perfect and up-to-date representation of any of the target populations applying for business surveys. Please try to indicate the nature and extent (by means of a percentage) of the frame imperfections in the sampling frame with respect to

a) undercoverage due to missing categories: *Only new enterprises*

<	*1*	%

coding: *hardly any, a few, a lot, (not known/blank)* [2, 1, 0, 0]

b) undercoverage due to other causes (e.g. change of principal activity, misclassifications): *Some undercoverage due to misclassifications*

	3	%

coding: *hardly any, a few, a lot, (not known/blank)* [2, 1, 0, 0]

c) overcoverage (e.g., dead units, changes in principal activity): *Some overcoverage caused by dead units and change of legal form*

| | *3* | %

coding: a few, a lot, (not known/blank) [1, 0 ,0]

Q13: Is any overcoverage of the sampling frame with respect to the target population detected

[x] before sampling? Please, indicate or estimate the percentage of sample units detected: *0* %

[x] among respondents? Please, indicate or estimate the percentage of sample units detected: *1 0* %

[x] during follow up of non respondents? Please, indicate or estimate the percentage of sample units detected: *2 0* %

[ ] No information available

coding: moderately, good, (not known/blank) [1, 2, 0]

Sample design

Q14: Please indicate which part of the target population (excluding threshold) is surveyed exhaustively and which part is examined by sampling?

Sampled strata: 6-10, 11-20; Surveyed exhaustively: 21-50, 51-100, >100

Q15: In case the survey is (partly) based on a sample, please specify

- the variables used for stratification: *number of employees, kind of activity*
- how the strata are constructed:

	101	*102*	*102*	*104*	*105*
2 - 5	1	2	3	4	5
6 - 10	6	7	8	9	10
11 - 20	11	12	13	14	15
21 - 50	16	17	18	19	20
51 - 100	21	22	23	24	25
> 100	26	27	28	29	30

- the allocation scheme employed (apart from oversampling)

  [x] optimal or Neyman allocation

  [ ] proportional

  [ ] x-optimal allocation (proportional to size)

  [ ] other, ...

- the (gross) sample size and the percentage sampled in each stratum: ....

	101	*102*	*102*	*104*	*105*
2 - 5	714	97	26	386	207
6 - 10	418	84	24	391	234
11 - 20	377	85	20	360	164
21 - 50	385	63	28	311	142
51 - 100	105	38	13	149	77
> 100	23	12	7	48	32

Q16: Give for each stratum (including the strata which are examined exhaustively) the population size.

	101	*102*	*102*	*104*	*105*
2 - 5	2857	388	102	1542	829
6 - 10	1266	254	72	1184	710
11 - 20	753	169	39	719	328
21 - 50	385	63	28	311	142
51 - 100	105	38	13	149	77
> 100	23	12	7	48	32

Q17: If applicable, indicate which groups are oversampled: *None.*

coding: yes , no, (not known/blank) [1, 0, 0]

For what reasons are these groups oversampled?

[ ] high nonresponse in specific subgroups

[ ] to permit generalizations for subgroups

[ ] reduction of overall variance

[ ] other, ...

Questionnaire design

Q18: Has the current questionnaire been changed over the past 5 years, or are any changes planned for the future? *Yes, two years ago*

Q19: Have any of the following methods been used for evaluating the quality of the current questionnaire, or will they be used for an upcoming redesign (multiple answers possible)?

[ ] No respondents involved

[x] Informal test interviews with respondents

[x] Contacts with repondents organizations

[x] Focus group discussions

[ ] Cognitive methods

[ ] Other, namely: ....

coding: none, to some extent , extensively, (not known/blank, not applicable)

[0, 1, 2, 0]

Q20: What kind of support is offered to businesses in completing the questionnaire (e.g., help desk, special instructions for new businesses)? *General instructions for all enterprises*

coding: none, limited, extensive, (not known/blank, not applicable) [0, 1, 2, 0]

Response burden

Q21: It is very important to get a good estimation of the response burden. Please indicate or estimate the average time a responding unit had to spend on participation in this survey?

	4	hr	*3*	*0*	min

Q22: Are any measures taken to monitor and control response burden?

[x] No

[ ] Yes. Please, describe: ....

Data collection and follow up procedures

Q23: What kind of data collection method is used? If more than one data collection mode is used specify or estimate the percentage of data collected by each mode.

	Percentage
[x] Mail	*50%*
[ ] Telephone, including CATI	
[ ] Interview (face-to-face), including CAPI	
[x] Electronic (tape, diskette or EDI)	*50%*
[ ] Secondary sources	
[ ] Other, namely:...	

coding: slightly, moderately, highly, (not known/blank) [2, 1, 0, 0]

Q24: What kind of follow-up strategy is applied? Please indicate the methods that are used to increase response (e.g., letter, telephone call, field workers) and describe when these methods are issued.

A letter after 10 weeks and a telephone call 5 weeks after first reminder

coding: none, limited, extensive, (not known/blank) [0, 1, 2, 0]

Q25: Please specify or estimate the number of units that were contacted on a 2nd occasion.

4 2 %	counted number
%	estimated number
[ ]	no information available

Q26: Please give the final response rates for each stratum?

	101	*102*	*102*	*104*	*105*
2 - 5	73	67	82	69	74
6 - 10	76	66	81	79	81
11 - 20	87	82	91	86	89
21 - 50	89	93	94	85	90
51 - 100	92	91	96	89	92
> 100	93	97	95	93	98

Data processing and estimation

Unit nonresponse

Q27: Please describe how much of the questionnaire has to be completed before it is considered a response case. *No information available*

Q28: How is unit non response dealt with?

[ ] Weight adjustment. Describe the methods used ....

[x] Imputation of all variables of interest. Describe the method used*: Replaced by average value of stratum*

[ ] Other: ....

coding (together with questions 41 and 42): none, simple, advanced, (not known/blank) [0, 1, 2, 0]

Q29: In case of imputation, is it possible to clearly identify imputed units in the 'clean' data file?

[x] Yes

[ ] No

<u>coding</u>: yes, no, (not known/blank*)* [1, 0, 0]

Q30: Is the impact of weight adjustment or imputation on the final estimates evaluated on a regular basis?

[x] Yes. How often are these evaluations carried out? *Annually*

[ ] No

<u>coding</u>: yes, no, (not known/blank) [1, 0, 0]

Item nonresponse

Q31: Please describe the methods of imputation used to replace missing values. ....

<u>coding</u>: simple, advanced, (not known/blank) [1, 2, 0]

Q32: Is it possible to clearly identify imputed values in the 'clean' data file?

[x] Yes.

[ ] No

<u>coding</u>: yes, no, (not known/blank) [1, 0, 0]

Q33: Please indicate or estimate the percentage fields imputed, for each stratum and key variable separately.

....

<u>coding</u>: small, considerable , (not known/blank) [1, 0, 0]

Q34: Is the impact of imputation on the final estimates evaluated on a regular basis?

[ ] Yes. Please describe the nature and extent of these evaluations: ....

[ ] No

<u>coding</u>: *yes, no , (not known/blank)* [1, 0 , 0]

Editing

Q35: The edit system used can be characterized as follows (multiple answers possible):

[x] Computer assisted interactive editing at the data collection or data entry stage

[x] Error lists generated by computer, followed by manual editing

[ ] Automatic data editing at the post entry stage

[ ] Plausibility inspection of aggregates, followed by a review of individual records underlying suspicious aggregates

[ ] Top-down editing by starting with the worst observation and stopping when further review has no impact on the survey estimates

[ ] Other, please describe: ....

coding (together with question 38): *micro, micro + aggregated, (not known, blank)* [1, 2, 0]

Q36: What kind of checks are used on individual records beside completeness checks (multiple answers possible)?

[ ] Routing checks

[x] Valid value checks

[x] Consistency checks

[x] Relational checks based on two or more questionnaire items

[x] Relational checks based on current and historical data

[ ] No checks at all

[ ] Other, please describe: ....

coding: none, limited, extensive, (not known/blank*)* [0, 1, 2, 0]

Q37: Please describe the methods used to identify records requiring manual review. *No special method used*

coding: none, simple, advanced, (not known/blank) [0, 1, 2, 0]

Q38: Please describe the methods used to adjust for errors. *Company contacted in case of errors*

coding (together with question 35): micro, micro + aggregated, (not known/blank) [1, 2, 0]

Q39: Is it possible to clearly identify adjustments in the 'clean' data file?

[x] Yes

[ ] No

coding: yes, no, (not known/blank)   [1, 0, 0]

Q40: Is the impact of editing on the final estimates evaluated on a regular basis?

[ ] Yes. Please describe the nature and extent of these evaluations. ....

[ ] No

coding: yes, no , (not known/blank)   [1, 0 , 0]

Estimation

Q41: What kind of estimation technique is used?

[ ] Horvitz-Thompson estimator

[x] Separate ratio estimator

[ ] Combined ratio estimator

[ ] Separate regression estimator

[ ] Combined regression estimator

[ ] Other...

coding: together with question 28

Q42: What kind of auxiliary information, if any, is used in the estimation process? *Number of employees*

coding: together with question 28

Eurostat/ statistical integration

Q43: To what extent do the survey data fit in with the data required by EUROSTAT, i.e., do the results of the survey process have the same degree of detail as asked by EUROSTAT? Please describe any differences. ....

Different size classes

Q44: Are any corrections made on the outcomes of the survey process for the purpose of integration before sending them to EUROSTAT?

[x] No

[ ] Yes. Please specify the differences between the survey estimates and the integrated estimates.
....

Costs

Q45: What is the most recent year for which you can give reliable estimates of costs as asked in the next questions? *1994*

Q46: How many persons are working in the department responsible for the industrial/business survey? *64*

Q47: How many of them are involved in conducting the survey in question? *40*

Q48: Can you give an estimate of the number of persons from other departments involved in the survey? *3*

Q49: It is important to get good estimates of the costs related to the survey activities. Therefore we ask you to complete the next table.

Q50: What are the total costs of personnel of the department responsible for the survey (in local currency)? *1.430.000*

Q51: What are the material costs, including costs for contract work, for the department carrying out the survey (in local currency)? *195.000*

Distribution of personnel in fte's over survey activities *)

	Number of full time equivalents involved in the survey on an annual base		
Activity	**From own department**	**From other department(s)**	**Total**
Use of sampling frame		*3*	*3*
Maintenance/update of the frame			
Quality inspection			
Drawing the sample			
Questionnaire design	*0,5*		*0,5*
Definition of questions			
Quality testing			
Data collection	*6*		*6*
Forms production			
Forms dispatch			
Forms receipt			
Data collection by field workers			
Follow up procedures	*2*		*2*
Response chase from inside			
Response chase by field workers			
Data entry	*10*		*10*
Data input			
Elementary validity checking			
Editing & imputation	*2,5*		*2,5*
Estimation, using external sources			
Statistical integration	*1*		*1*
Other survey activities	*3*		*3*
Analysis			
Disclosure checking			
Publication			
Inquiries			
Others			
Total	*25*	*3*	*28*

*) general management and computer support staff are <u>not</u> to be included

Glossary

These definitions are largely adapted from the work of the EUROSTAT working group on "Assessment of the Quality of Business Statistics " (EUROSTAT, 1996b).

ACCOUNT MANAGEMENT

Account management (or profiling) involves personal contact with (usually) large businesses to gain insight into their legal and organisational structure, and to define the statistical structure for the business register.

ALLOCATION

The way in which the total sample size is allocated to each of the strata in a stratified design.

- *Optimal or Neyman allocation:* the total sample size is allocated in such a way so as to maximise precision subject to fixed total costs

- *X-optimal allocation:* optimal allocation (see above) using an auxiliary variable that is highly correlated with the target variable.

- *Proportional:* allocation proportional to stratum population size.

For details see, for example, Särndal, Swensson, and Wretman (1992).

BUSINESS REGISTER

Multipurpose frame database designed to provide frames for individual business surveys. It should be an up-to-date file of all statistical units from which the frames are extracted. Registers are generally derived from administrative data.

CATI

Computer-assisted telephone interviews

CAPI

Computer-assisted personal interviews

COVERAGE ERRORS

Coverage errors may involve either the failure to identify and count units which should have been included or the erroneous inclusion in the sample of units not properly part of a census or survey.

DUPLICATES

Within a list or frame, a duplicate is a record that represents the same business entity as another record in the same list.

EDITING

Editing is the procedure for detecting and adjusting individual errors in data records resulting from data collection and capture. The checks for identifying missing, erroneous, or suspicious values in computer-assisted editing are called edit rules or edits. An editing change occurs when an item (question) value is adjusted as a consequence of action taken when an error is identified. Micro-editing and macro-editing may be distinguished in order to calculate rate of edits.

ELECTRONIC DATA INTERCHANGE (EDI)

EDI is the electronic transfer of business transaction information in a standard format between business partners. EDI offers businesses the opportunity to retrieve information electronically from their internal systems and to forward that information to trade partners/ suppliers/ customers/ government through a communications network.

ERROR

The difference between the survey value and the corresponding true value.

ESTIMATE

An estimate is the calculated value using an estimator.

ESTIMATOR

An estimator is the mathematical function by means of which the estimate for a particular parameter is computed.

FOCUS GROUP

A group of comparable respondents who participate in a structured group discussion (often using audio/video registration) in order to evaluate questions and instructions, and to identify errors or burdens associated with understanding and answering questions and retrieving data.

FOLLOW-UP

A procedure whereby those members of a selected sample for whom a response is not obtained by one data collection strategy (e.g., telephone or mail) are contacted by the same or another data collection strategy in order to increase response rate. It can also be used to designate repeated surveys among a panel of respondents.

FRAME

The frame for a survey is the listing or listings of units in the study population from which the sample is drawn and through which units of the study population are contacted.

FRAME IMPERFECTIONS

Frame imperfections may be caused by the inherent limitations of input data, or by delays and errors in the data acquisition and processing. Frame errors cover:

- coverage errors - erroneous inclusions, omissions and duplications;

- classification errors **-** units not classified, or misclassified by industry, geography, or size;
- contact errors **-** units with incomplete or incorrect contact data**.**

IMPUTATION

Imputation refers to the replacement of missing data using logical edits or statistical procedures. It concerns item nonresponse or unit nonresponse.

ITEM NONRESPONSE

See non response error.

MACRO-EDITING

A procedure for pointing out suspicious data by applying checks/edits based on weighted keyed-in data. This method can reduce the manual verifying work of suspected data by 35-80% as compared to corresponding traditional micro-editing methods without any loss in quality.

MICRO-EDITING

Editing which is applied on all the questionnaires as opposed to macro-editing.

MISCLASSIFICATION

Misclassification refers to wrongly classified units, for example by industry, geography or size.

NONRESPONDENT

Enterprises in a sample from which no information has been obtained.

NON RESPONSE

The failure to obtain complete information for all units in the selected sample.

NON RESPONSE ERROR

Non response error results from a failure to collect complete information an all units in the selected sample except non eligible units (dead units, out of scope units, non contacts). There are two types of nonresponse error. First, a sampled unit that is contacted may fail to respond. This represents unit non response. Second, the unit may respond to the questionnaire incompletely. This is referred to as item non response.

OVERCOVERAGE

Error due to the units which are included in the sampling frame but do not belong to the target population. These cases are usually observed for contacted units, but not necessarily for noncontacted units or those excluded from a sample. Reasons for overcoverage are death of units, misclassification and a non-updated frame.

QUALITY SURVEY, QUALITY INSPECTION

Periodically conducted survey especially aimed at updating the register to ensure it is consistent and accurate at a given point in time.

RESPONSE RATE

The percentage of an eligible sample for which information is obtained. For an interview survey the numerator of the formula is the number of interviews. The denominator is the total sample size minus non-eligible respondents; that is, minus those not meeting the criteria for a potential respondent as defined for that particular study.

SAMPLING DESIGN

The sample design should provide information on the target and final sample sizes, strata definitions and the sample selection methodology. It should also include details of changes in methodology and any impact these changes have on comparability over time.

STUDY POPULATION

This is the actual population at which the survey is aimed, i.e., the set of units from which the survey sample is actually selected, and from which estimates for the target population will be derived.

TARGET POPULATION

This is the conceptual population of units at which the survey is aimed and about which information is wanted. For practical reasons this population of units often cannot be assembled exactly. The target population will generally be different from the population actually sampled (i.e., the study population).

UNDERCOVERAGE

A type of non sampling error that results from failure to include in the frame all units belonging to the defined population. This mainly includes (new) enterprises not included in the frame, either through real birth or demergers, and to misclassified units.

UNIT NONRESPONSE

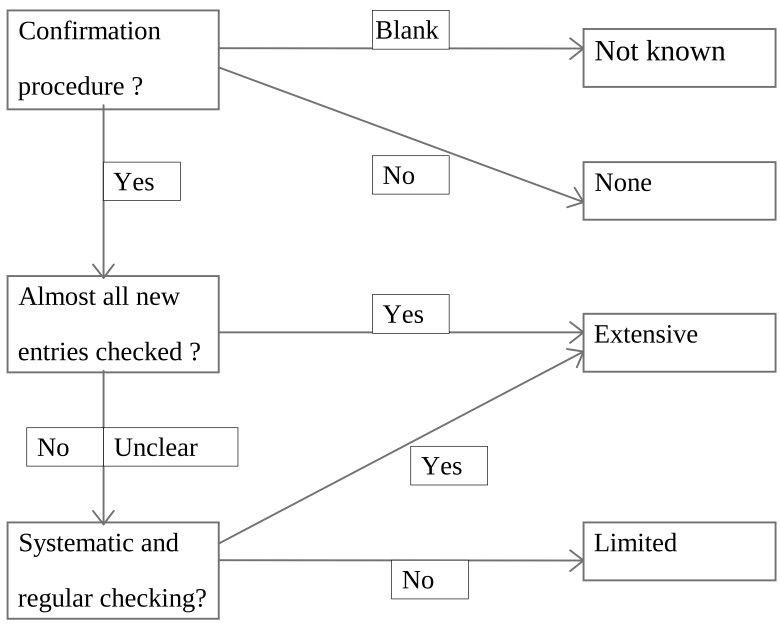
See nonresponse error.

Appendix E: Coding the questions from the business questionnaire

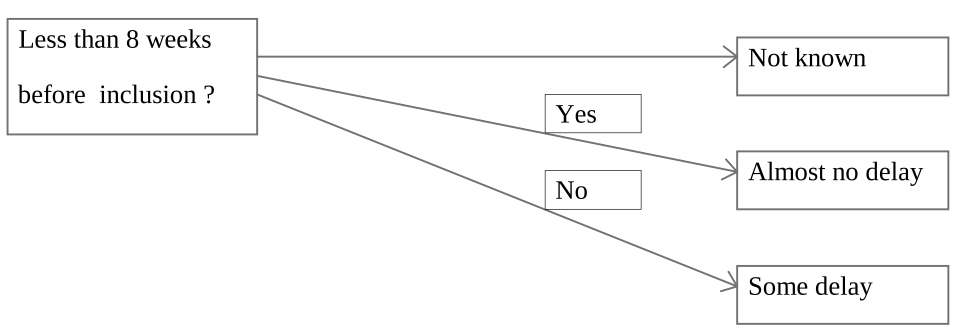
This appendix indicates how the questions from the business questionnaire can be translated into a set of pre-defined answer categories. For the questions than cannot be coded straightforwardly, one or more decision rules are given in order to arrive at an answer category. It should be noted that the criteria used in these decision rules are not to be considered as absolute standards. They are merely suggestions on our part for what we believe constitute reasonable boundaries for coding the questions. Furthermore, the answers from completed questionnaires should give an indication as to whether or not these rules can be sensibly used in practice.

Further note that not all the questions from the questionnaire are coded into pre-defined answer categories. Some questions are used only for background information on the statistical process. In case a question is left unanswered or the answer is not clear it is coded as ‘Not known’.

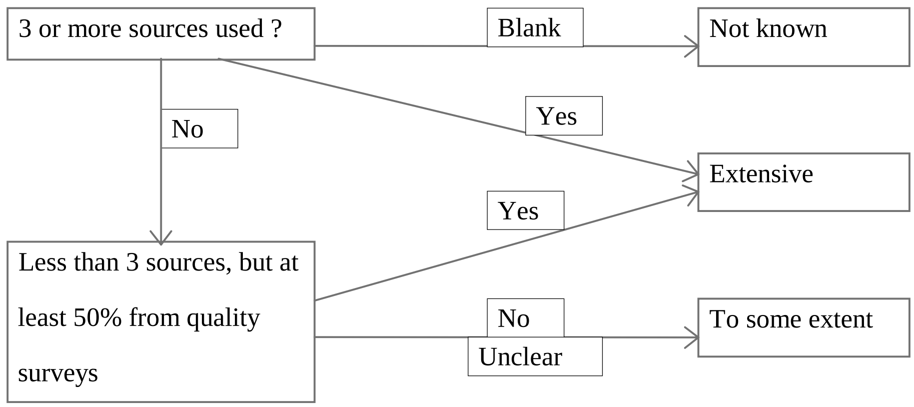
Q6: Is any conformation procedure carried out before potential births of populations units are recorded on the register? *Process characteristic*: inspection of births.



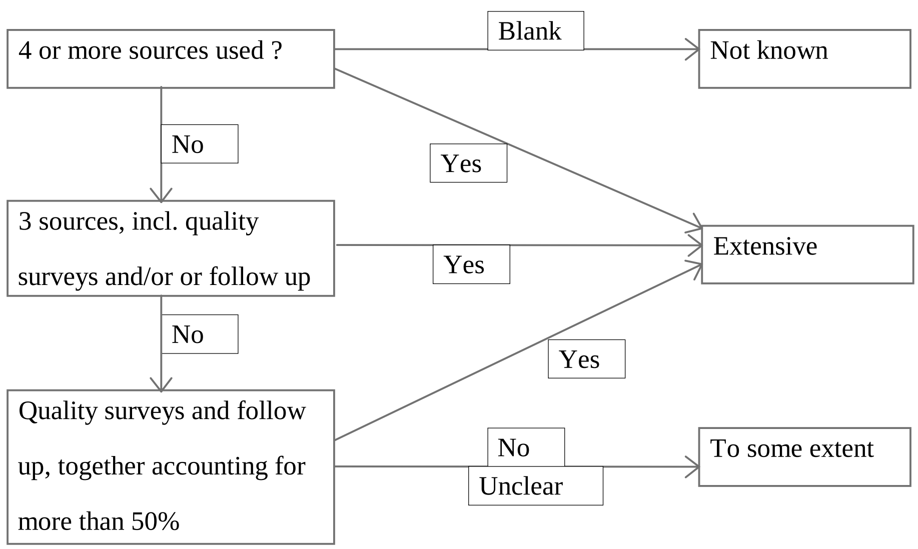
Q7: How long on average is the delay between birth and inclusion of a population unit in the business register? *Process characteristic*: timeliness registration of births.



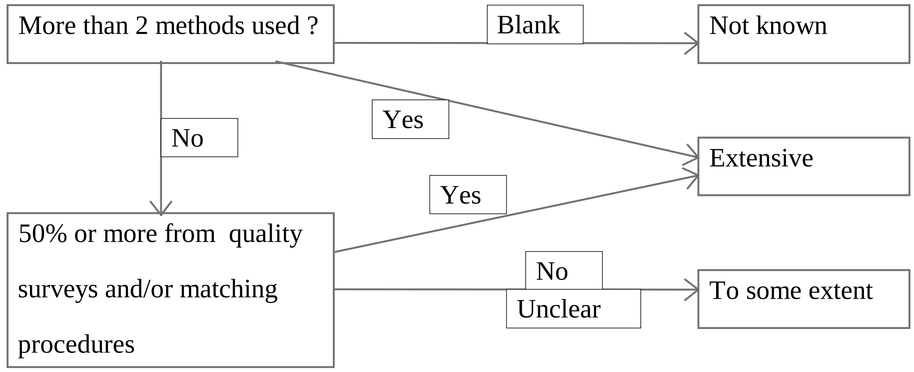
Q8: Indicate what kind of sources are used to update names, addresses, size indicators, classification codes and the structure of business in the register? *Process characteristic*: correction for misclassifications.



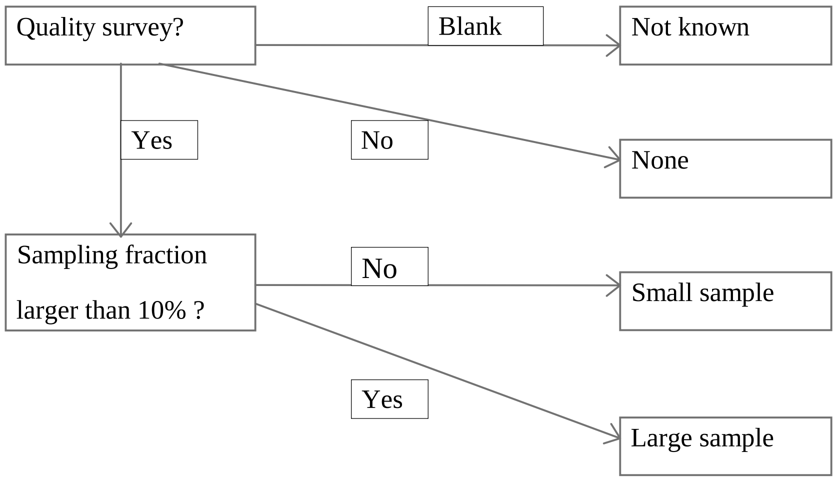
Q9: Indicate what kind of sources are used to identify ‘dead’ population units in the business register? *Process characteristic*: correction for deaths.



Q10: Indicate how duplicate units in the business register are identified? *Process characteristic*: correction for duplicates.

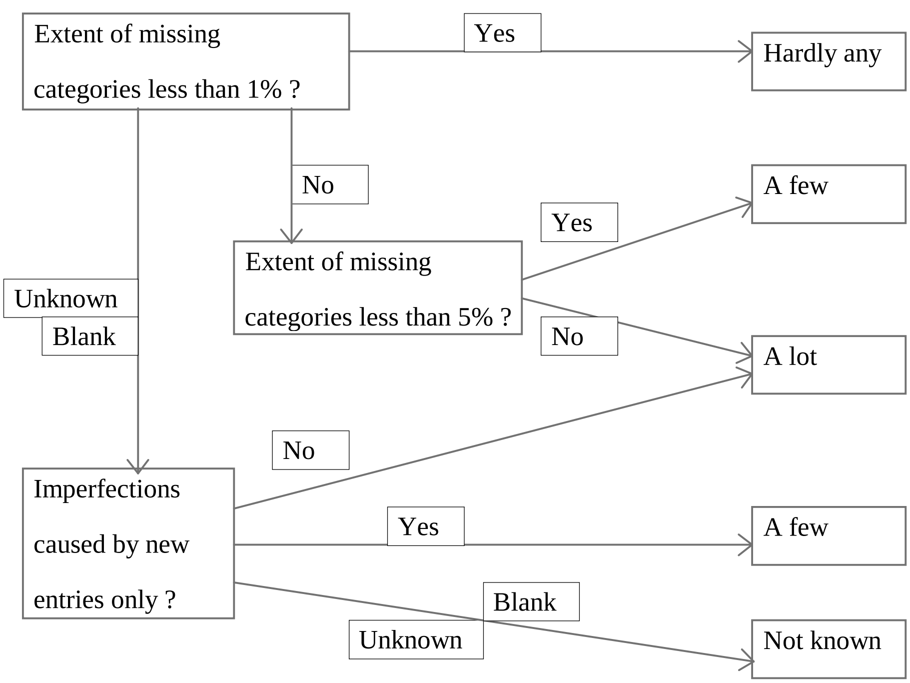


Q11: **a**. If special quality surveys are carried out to maintain the register, specify the sampling fraction. *Process characteristic*: size of quality survey.

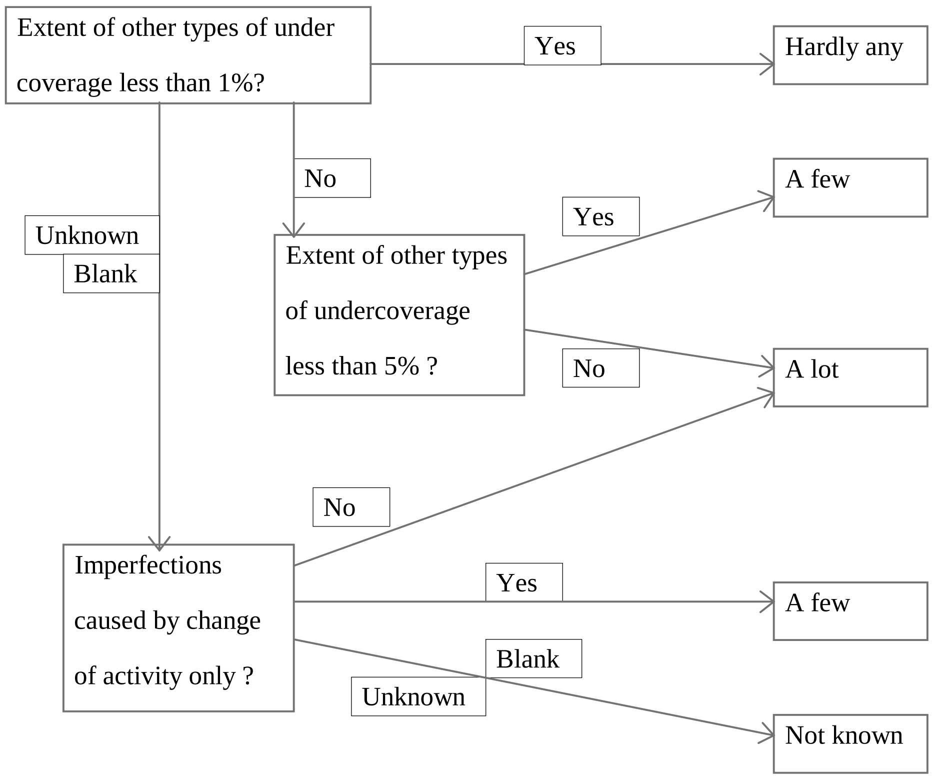


Q11: **b**. If special quality surveys are carried out to maintain the register, specify the frequency with which these samples are taken. *Process characteristic*: frequency of quality survey.

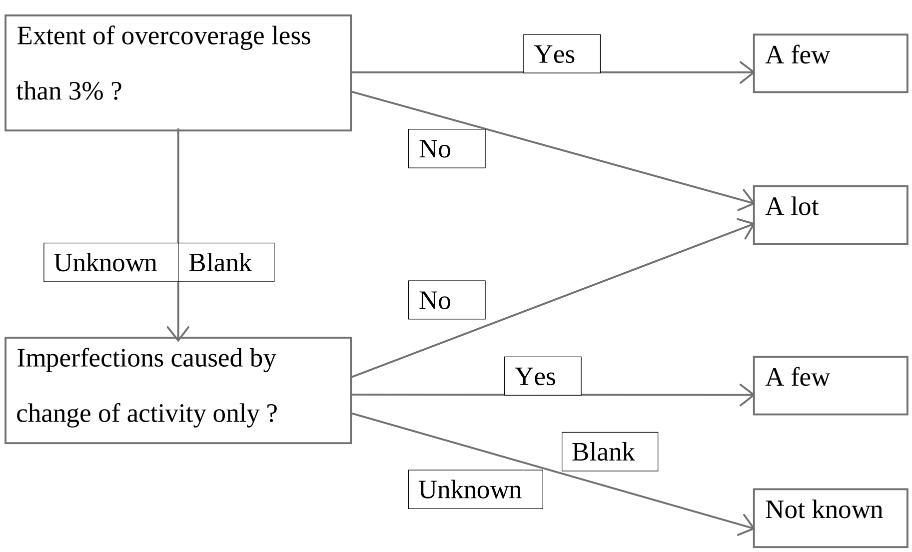
Q12: **a**. Try to indicate the nature and extent of the frame imperfections in the sampling frame with respect to undercoverage due to missing categories. *Process characteristic*: missing categories.



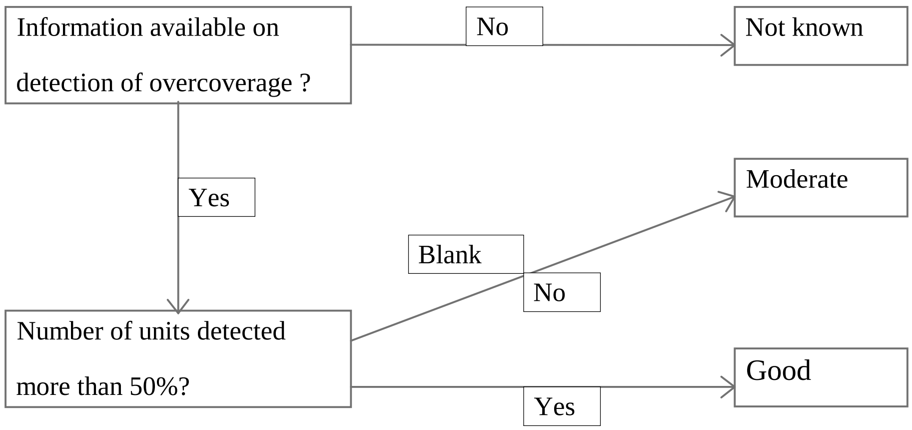
Q12: **b**. Try to indicate the nature and extent of the frame imperfections in the sampling frame with respect to undercoverage due to other causes (change of activity, misclassifications). *Process characteristic*: <u>other types of undercoverage</u>.



Q12: **c**. Try to indicate the nature and extent of the frame imperfections in the sampling frame with respect to overcoverage (dead units, change of activity). *Process characteristic*: <u>dead or misclassified units</u>.

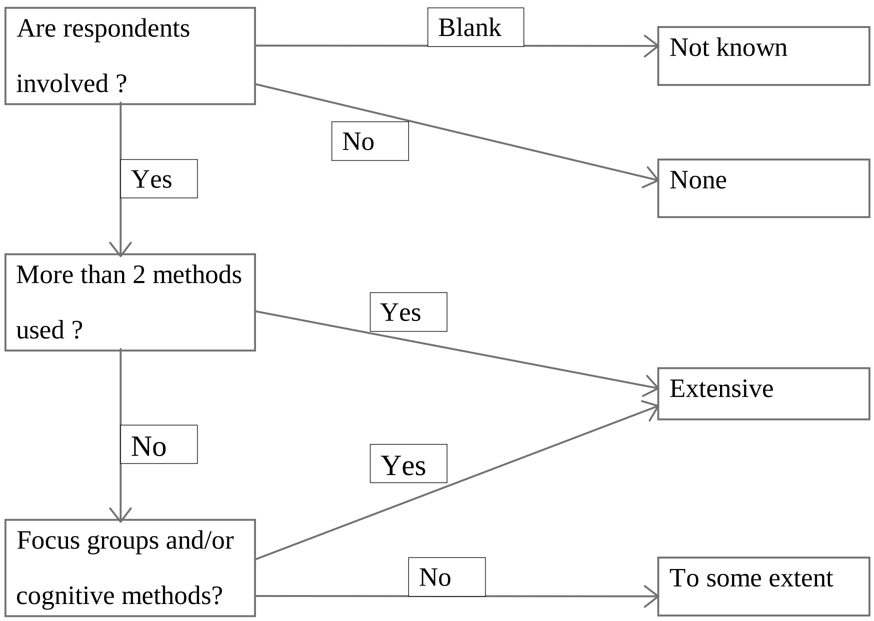


Q13: Is any overcoverage of the sampling frame with respect to the target population detected? *Process characteristic*: detection of overcoverage.

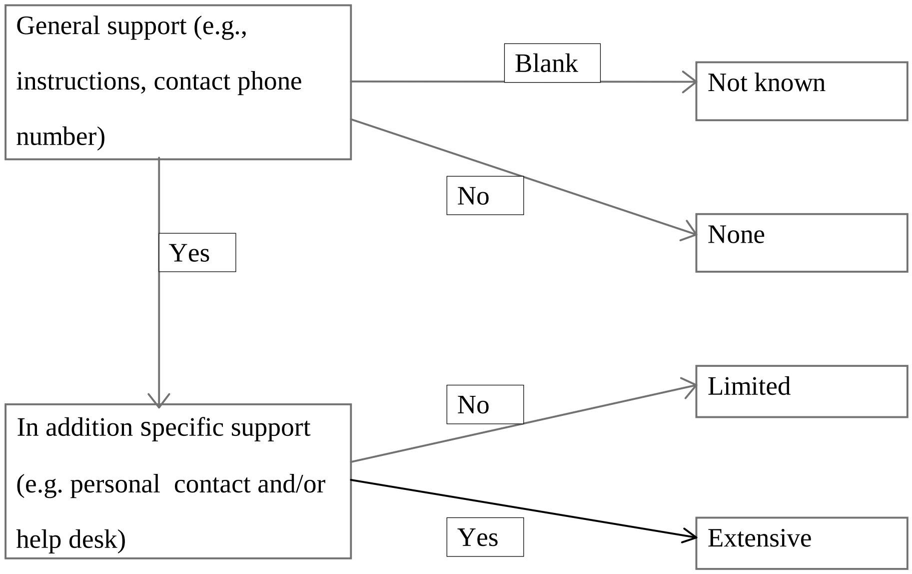


Q17: If applicable, indicate which groups are oversampled. *Process characteristic*: oversampling for non response.

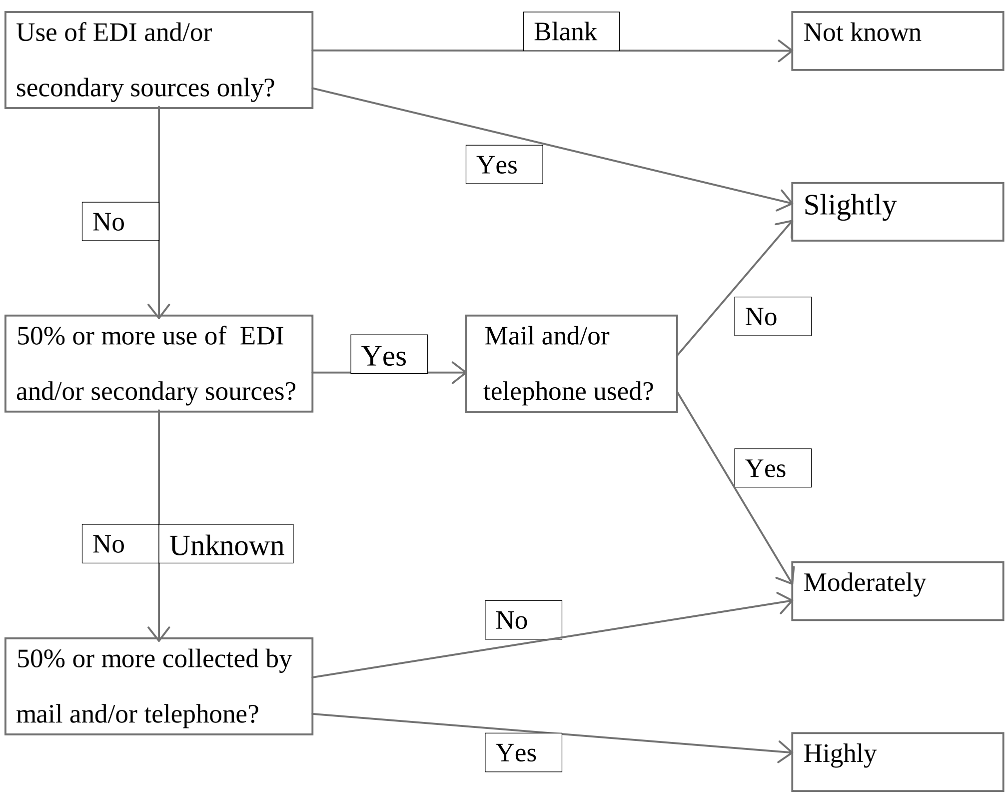
Q19: Have any of the following methods been used for evaluating the quality of the current questionnaire, or will they be used for an upcoming redesign? *Process characteristic*: participation of respondents.



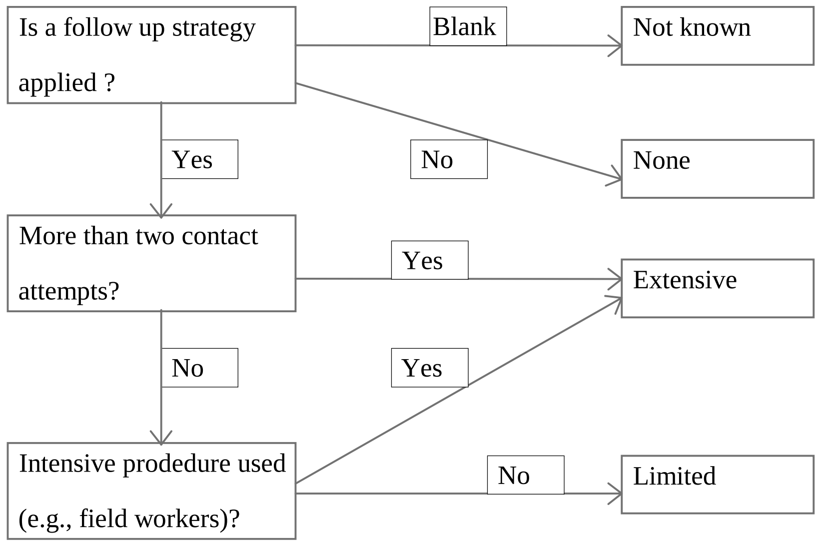
Q20: What kind of support is offered to businesses in completing the questionnaire? *Process characteristic*: support in completing questionnaire.



Q23: What kind of data collection method is used? *Process characteristic*: non response sensitivity of data collection method.



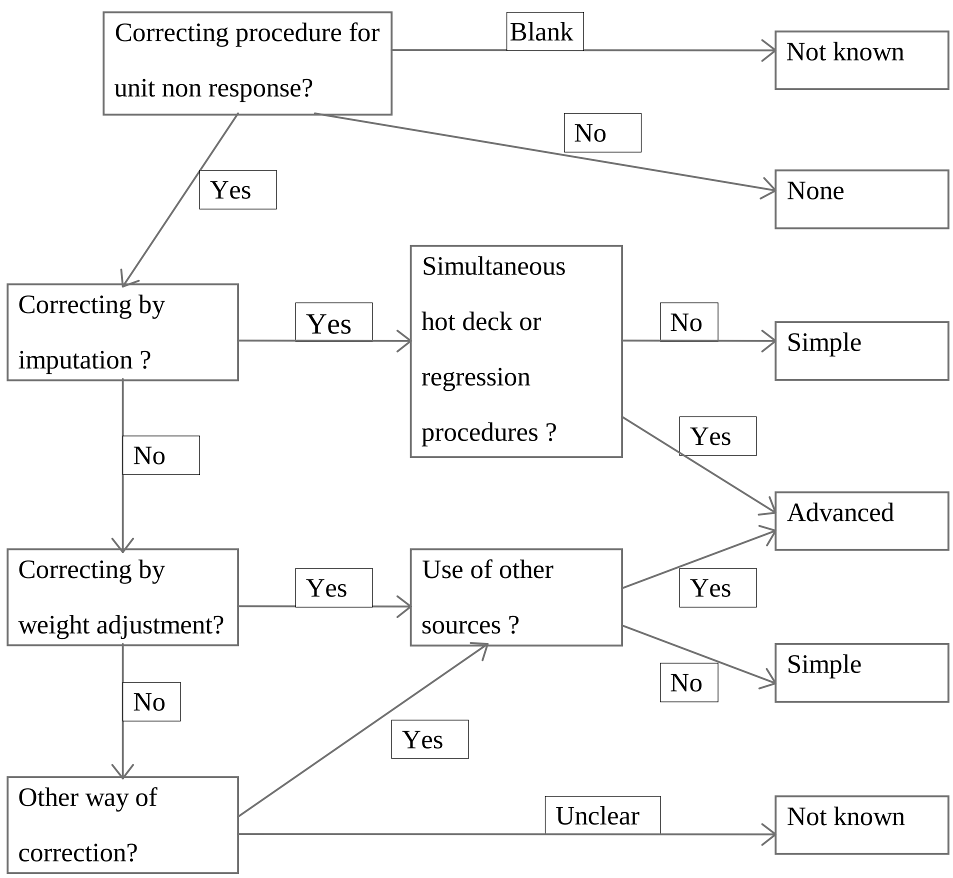
Q24: What kind of follow-up strategy is applied? Indicate the methods that are used and describe when these methods are used? *Process characteristic*: intensity follow up procedure.



Q28: How is unit non response dealt with?

Q41: What kind of estimation technique is used?

Q42: What kind of auxiliary information, if any, is used in the estimation process? *Process characteristic*: methods to correct for non response.



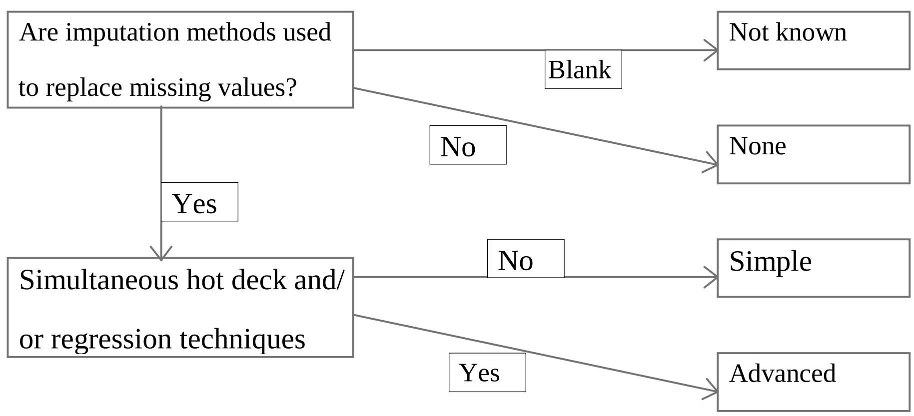
Q29: In case of imputation, is it possible to clearly identify imputed units in the 'clean' data file? *Process characteristic*: recognizing imputed units.

Yes	No	Not known

Q30: Is the impact of weight adjustment or imputation on the final estimates evaluated on a regular basis? *Process characteristic*: evaluating impact of weight adjustment/imputation on estimates.

Yes	No	Not known

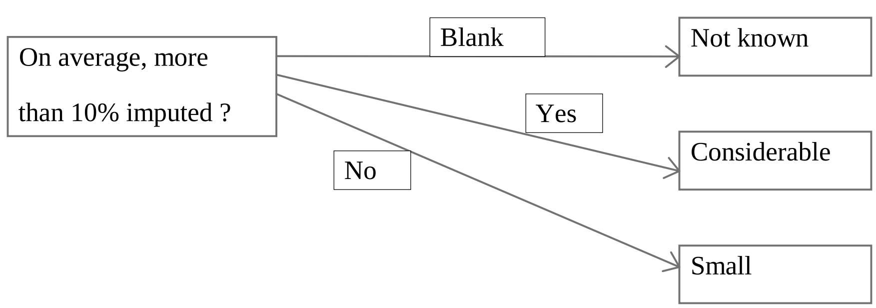
Q31: Describe the methods of imputation used to replace missing values. *Process characteristic:* qualification of imputation method.



Q32: Is it possible to clearly identify imputed values in the 'clean' data file? *Process characteristic:* recognizing imputed items.

Yes	No	Not known

Q33: Indicate or estimate the percentage fields imputed, for each stratum key variable separately. *Process characteristic*: extent of imputations.

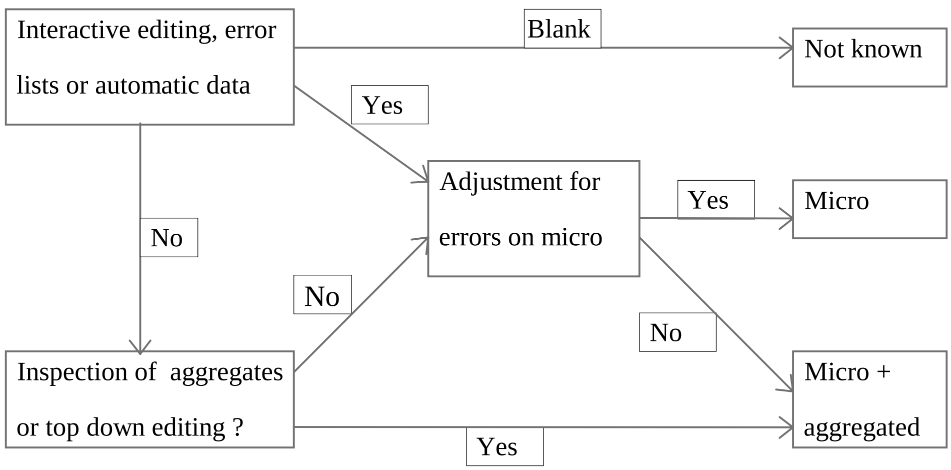


Q34: Is the impact of imputation on the final estimates evaluated on a regular basis? P*rocess characteristic*: evaluating impact of imputations on estimates.

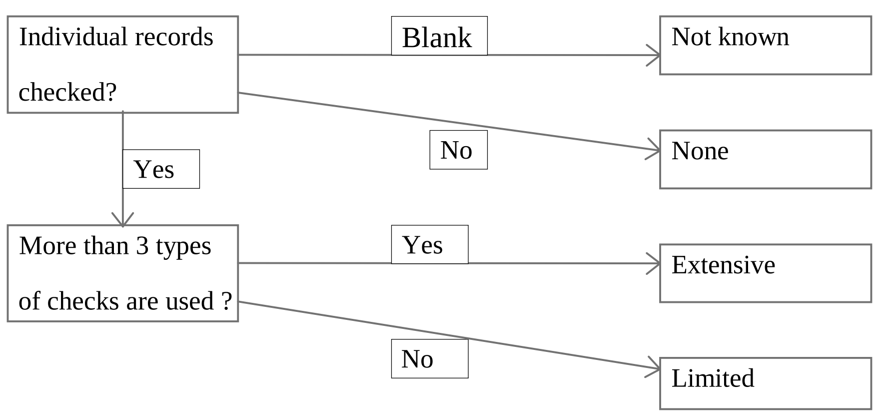
Yes	No	Not known

Q35: The edit system used can be characterized as follows.

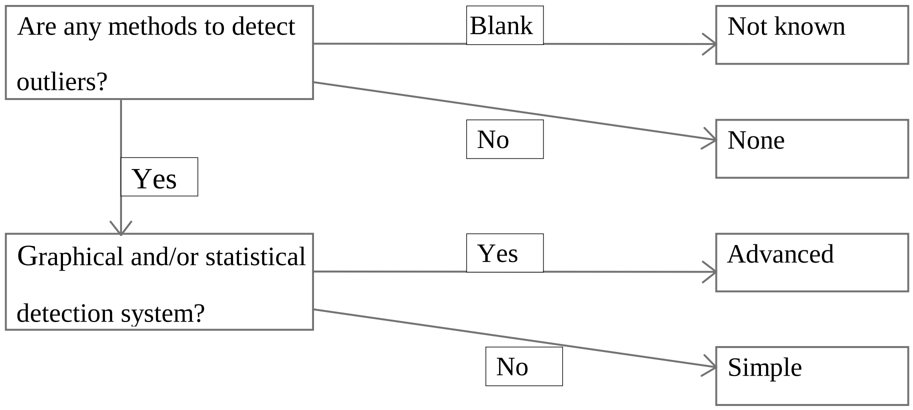
Q38: Describe the methods used to adjust for errors. P*rocess characteristic*: qualification of edit system.



Q36: What kind of checks are used on individual records beside completeness checks? *Process characteristic*: intensity of checks on records.



Q37: Describe the methods used to identify records requiring manual review. *Process characteristic*: detection of records requiring manual review.



Q39: Is it possible to clearly identify adjustments in the 'clean' data file*? Process characteristic*: recognizing edits.

Yes No Not known

Q40: Is the impact of editing on the final estimates evaluated on regular basis? *Process characteristic*: evaluating impact of editing on estimates.

Yes No Not known

References

Bethlehem, J.G, and H.M.P. Kersten, 1985, *On the treatment of non response in sample surevyes*. Journal of Official Statistics 1, pp. 287-300.

Cochran, W.G., 1977, *Sampling techniques*. John Wiley & Sons.

EUROSTAT, 1996a, “*Proposal for a standard quality report on business statistics*”, Eurostat/D3/Quality/96/04.

EUROSTAT, 1996b, “*Glossary on quality of business statistics*”, Eurostat/D3/Quality/96/03.

EUROSTAT, 1997, “*Quality in business statistics*”, Eurostat/A4/Quality/97/02.

EUROSTAT, 1998, A.J. Willeboordse (eds.), *Handbook on the design and implementation of business surveys*.

Lessler, J.T and W.D. Kalsbeek, 1992, *Non sampling error in surveys*. John Wiley & Sons.

Särndal, C-E., B. Swensson, and J. Wretman, 1992, *Model assisted survey sampling*. New York, Springer Verlag.

Statistics Netherlands, 1995, *Proposal for Lot 13: Data Quality* (Statistics Netherlands).

Willeboordse, A.J., 1995, *Monitoring the response burden for business surveys*. Internal report (Statistics Netherlands).