Netherlands Official Statistics

Volume 15, Summer 2000

Special issue

Integrating administrative registers and household surveys

Voorburg

Prinses Beatrixlaan 428
P.O. Box 4000
2270 JM Voorburg (Netherlands)

Telephone : . .31 (070) 337 38 00
Fax : . .31 (070) 387 74 29
E-mail: lhka@cbs.nl
Internet: http://www.cbs.nl

Heerlen

Kloosterweg 1
P.O. Box 4481
6401 CZ Heerlen (Netherlands)

Telephone : . .31 (045) 570 60 00
Fax : . .31 (045) 572 74 40

Key figure A-125/2000

Subscription: Dfl. 42.00 per year
Price per copy: Dfl. 20.00

ISSN 0920-2048

Postage will be charged.

Contents

Re-engineering social statistics by micro-integration of different sources: an introduction

Pieter G. Al and Bart F.M. Bakker

1. Introduction

In the last three years, Statistics Netherlands has developed a statistical database system that contains almost all of the relevant information it holds on persons, households and quality of life. The database has been created by matching records from the available administrative registers and the household sample surveys. The development of the database was dictated by changing demands for statistical information and increasing non-response rates in household sample surveys. It was made possible by the recent rapid developments in information technology and a changing attitude in the Netherlands towards confidentiality aspects of data linking. This introductory article discusses the background of these developments and gives an overview of the articles in this special issue.

2. Changing demands for statistical information

The increasing prosperity following World War II shifted the political focus from the mere growth of prosperity to its distribution, prompting the introduction of many subjects to the political agenda that had not been there previously. There was a growing interest in subjects like poverty and social exclusion, and the social situation of smaller groups in society increasingly received attention.

This led to drastic changes in policy-makers' demands for social statistics in recent decades. Monitoring social developments has become an important element of modern policy-making, and users need data which are relevant and authoritative. Users want time series for the most relevant indicators on each policy issue, and these should be undisputed because of their legitimising function. In order to be authoritative, statistics on important social phenomena should be consistent with each other.

The depth of the statistical information policy-makers need has also changed: they no longer want only 'one-dimensional' statistics on main political issues like health, crime, education, employment and income. There is also an increasing demand for information on the usually complex relationships between these themes. This demand reflects the policy issues arising from societal developments such as poverty, social exclusion, unemployment and deprivation, problems that often converge in rather small groups in society or in certain areas like inner cities. To fulfil the politicians' requirements statisticians have to compile coherent sets of indicators providing information on the accumulation of effects that have recently become a policy issue.

Important policy issues have a bearing on specific groups in society, ethnic minorities for example. To be able to provide information on all aspects of such issues, data from various administrative registers and household sample surveys have to be combined. Another important development is the demand for information on phenomena with a small incidence, for example victims of certain crimes or incidence of specific diseases.

In short, users want relevant and authoritative statistical information, which gives insight into the complex relationships between different aspects of social and economic life. This information has to be detailed enough to indicate the situation of small groups in society and to estimate phenomena with a small incidence. The information should also be made available early in order to describe important developments in society.

3. Changes in the production of statistical information

In the past, economic growth itself as source of prosperity received much more attention than the subsequent distribution of this prosperity. Because of this, economic statistics were developed much earlier than social statistics. A lot of effort was put into compiling a system of national accounts, for example, which are now nationally and internationally widely recognised as relevant and authoritative statistics whose validity is seldom questioned.

There is no such uniform theoretical base for the topics to be covered by social statistics, however. For some areas in social statistics accounting systems similar to the system of national accounts have been developed. Labour accounts, for example, cover the labour market, and socio-economic accounts cover the field of household income and expenditure. Although each of these systems provides consistent data for its specific field, there is no consistency of data for the whole field of social statistics. In addition to the partial systems, at present theoretical approaches for social statistics are manifold and consequently existing lists of social indicators vary greatly in background, size and content. There is only a limited international consensus on 'core indicators' in the field of social statistics: only for some domains of social statistics and only in some regions of the world. To a certain extent, the described increasing demands for information are related to lack of a sound framework.

In line with statistical tradition, one way to fulfil the new demands for statistical information is to develop household sample surveys that are ever more comprehensive. These surveys aim to cover as many variables as possible, so that all relevant relations can be analysed in a comprehensive data set. The sample size should also be large enough to allow description of relatively small subgroups and relatively rare phenomena.

However, there is a growing awareness of the limits to this approach. The response burden on the sampled households becomes heavier when more variables are covered in a single survey and this constitutes a serious constraint on statistical agencies to meet the user demands. Moreover, there are the high non-response rates for household sample surveys, which render survey estimates questionable because of a potential bias that is hard to measure. Non-response rates for Statistics Netherlands' sample surveys have increased sharply in the last decades, prompting a growing concern about the problem of bias in recent years. The 90% response to the Labour Force Survey (LFS) in 1977 had dropped to only 60% by 1995. For the Quality-of-life surveys, the response rates fell from 72% in 1974 to 50% in 1995. Most non-response – some 65% – consists of refusals and is concentrated among singles, young adults, ethnic minorities and residents of cities, making it highly selective. An international study on sample survey non-response showed that Dutch non-response rates are very high compared with other countries.

And lastly, sample surveys among households do not come cheap: people are less and less inclined to participate in surveys making it more and more expensive to gather a sufficient number of records. This leads to serious budgetary problems for the statistical office at a time that budgets are already under sever pressure.

4. The use of information from various registers

The last conventional population census in the Netherlands was conducted in 1971. During the seventies there was increasing concern about the protection of privacy in the Netherlands. Although the 1971 census prompted the public debate on the subject, it affected the response rate for this census only slightly: ultimately non-response was 0.026%. However, the rapidly growing concern caused a postponement of the 1981 census because of the estimated risk of non-response as high as 26% of the population, and later the decision was taken to abandon conventional censuses altogether because of the fear of very low response rates.

But the demand for census information did not disappear, and Statistics Netherlands had to find alternative ways to supply this information. One possibility to provide information on the population is to use all kinds of sources, be they household sample surveys or administrative registers, and choose the best source for each part of the information required. Although the Dutch census programmes of 1981 and 1991 provided the information in this way, they contained inconsistencies at a macro-level. One important reason for this was that register data were only available in an aggregate form (for example: social security benefits) and the outcomes could not always be correctly integrated with outcomes from household sample surveys.

In the course of the years, the uproar about the protection of private data brought about by the 1971 census shifted into a public debate and subsequently into a process of legislation, culminating in the Act on Registrations of Personal Data (WPR) in 1988. This law regulates the maintenance and use of registrations of personal data, overseen by the Registration Chamber; it is soon to be replaced by the new Personal Data Protection Act (WBP), which incorporates special provisions for data needed for historical, statistical and scientific purposes. In cases where data are used solely for research in these fields of research, and security is guaranteed, the act creates wider possibilities for data use than the original purpose for which they were obtained. In addition, they may be kept for longer. So the early concerns about personal privacy have resulted in well-balanced legislation. In the end, this creates new possibilities for the national statistical office to use data from administrative registers.

Statistics Netherlands warmly welcomed this step: administrative registers constitute a new source of information, and one that is relatively cheap to use. Large numbers of records at a time can be obtained from institutions like the population registration, the social security institutions and others. Moreover, the possibilities of electronic data handling and electronic data interchange have grown rapidly in the past decades, and for social statistics have proven to be very stimulating to gather information for example from employers on their employees. Data on job characteristics, hours worked, wages and social security premiums paid are now sent to Statistics Netherlands by many enterprises or their external wage administrators. Not only is this an efficient way to gather a lot of valuable information, it also substantially reduces the response burden for companies and institutions. So it is now policy at Statistics Netherlands to make optimum use of administrative register information in order to reduce respondent burden and costs.

5. Micro and macro integration of different sources: the SSB

Our long experience of national accounting and the partial accounting systems developed for social statistics has taught us that integrating data from different sources leads to great improvements in the quality and scope of statistical data. Until now, the accounting systems have been produced on a rather aggregated level, with tables compiled on the level of intended publication. A drawback of this practice is a loss of flexibility, making it difficult to produce tables with different dimensions. Moreover, the many different aspects of social life complicate the field of social statistics much more that that of economic statistics, making it much harder to describe social statistics in a system.

The administrative registers often contain complete information on all relevant units. In the Netherlands, this is certainly the case for demographic data, labour market participation, dependence on social security benefits, participation in education and housing facilities. There are also other aspects for which complete registers are available but have not until now been tested by Statistics Netherlands.

The completeness of the information in the registers soon prompted proposals for micro integration, and Statistics Netherlands started a research programme to explore the possibilities of compiling statistics on persons and households by matching, editing, integrating, imputing and weighting data from administrative registers and household sample surveys combined. This research programme is called the *Sociaal Statistisch Bestand* or SSB in Dutch , which translates as Social Statistical Database. The Population Register files form the backbone of the database, as all the other files are linked to this register. The first phase of this programme consisted of a pilot study, and following its success in the second phase a prototype of the SSB was built with available data for 1995. In the third phase, the aim is to build a prototype in which all the available data for 1995-1998 are matched.

One of the main basic ideas of the data input in the SSB is the integration of data from different sources. When records from different sources are matched, there is a check for the completeness of the registers and the occurrence of double records. This can already be seen as a form of micro-integration. After linking, the statistical variables have to be derived from the characteristics recorded in the linked records. Using all the information available about one and the same unit is another form of micro integration. This kind of data editing leads to consistent statistical information for all variables processed, and also makes later macro-integration easier as many inconsistencies are removed at the beginning of the process. For missing data for smaller fractions of the total population, imputation can be used as a form of integration to arrive at completeness. For example, if paper boys not recorded under jobs, these jobs have to be added to the register using information on the number of newspapers delivered. Moreover, these jobs have to be assigned to people in the various regions of the country. Here macro and micro-integration meet each other. Of course, after these micro integration stages, there is also macro integration: for example, the total expenditure of government on social benefits must equal the total amount of received benefits at an individual level.

Ultimately, the SSB will contain all the available relevant information. For aspects for which only information from sample surveys is available, this database contains the maximum amount of information there is to raise the survey data. This is all the more important in view of the low response rates in the Netherlands. The database provides an insight into the complex relationships between different aspects of social and economic life, and at the same time contains maximum detail so that the situation of small groups in society can be analysed and phenomena with a small incidence can be estimated. The information should be produced yearly in order to describe important developments in society.

6. Contents of this special issue

The design and organisation of the statistical process is radically changing at Statistics Netherlands. The change has been triggered by the growing demand for coherent statistical information, by developments in information and communication technology, by high non-response rates in household sample surveys and by

political pressures to cut down staff costs and to minimise the reporting burden caused by statistics. The article by Van der Laan gives an overview of the design of the SSB and the underlying statistical processes. He focuses on the integration of administrative registers and household sample surveys at the micro level as an essential part of the redesigned statistical process, demonstrating that this approach to social statistics will improve the coherence in the statistical output.

In their article, Arts, Bakker and Van Lith describe the matching of the sources, both administrative registers and household sample surveys. Matching on a personal identification number proves successful: between 96 and 98 percent of the records match. In the absence of such identification numbers, the sources are matched on postal code, house number, date of birth and sex. This results in 93 to 95 percent of matching records. A method to estimate the expected number of false matches is found to be of very limited use. In order to improve the linking procedure, samples are taken from one of the population registers.

Once the administrative registers and sample surveys have been linked, it is necessary to estimate the frequencies and cross tabulations to be published. The article by Kroese, Renssen and Trijssenaar describes the tests of several alternative estimation strategies to be used in the new statistical process of the SSB. The first strategy of repeated weighting can be seen as a new application of old weighting techniques. The second strategy is the method of mass imputation. Some preliminary results of a project are presented in which the methods are tested. It is argued that the method of mass imputation is less attractive than the method of repeated weighting.

Non-response rates are particularly high in the Netherlands and render sample survey estimates questionable, as they introduce a potential bias that is difficult to measure. One of the possibilities of the SSB is to correct more effectively for selective non-response. Matching the administrative registers with sample survey information makes it possible to search for the characteristics that correlate with the probability of response and with the target variables in the sample survey. Furthermore, it is possible to select those characteristics to weight the sample survey. Geuzinge, Bakker and Van Rooijen present an empirical example in which the health interview survey is weighted with the use of administrative registers on jobs and social security benefits. They show that the developed method has serious advantages over traditional methods for weighting the data.

One of the advantages of the SSB approach is that a large part of the integration process that up to now had been done at the macro-level will be performed at a micro-level. However, macro-integration will still be necessary, for example to ensure consistency with the national accounts or with the outcomes of business surveys. The article by Bakker and Van Rooijen describes the macro-integration process of business and household sample surveys on the quantities of services within the frame of the so-called supply-and-demand matrices. In this integration process several adjustment factors are estimated. After the determination of the integrated quantities of services, it is necessary to attain consistency between the macro-totals and the aggregated micro-totals in the SSB.

As the SSB will ultimately comprise a very detailed picture of each inhabitant of the Netherlands, data security and confidentiality are very important issues. Statistics Netherlands cannot run the risk of individual data being disclosed because if they are, the support of the Dutch population and the government will be nil. Beside this consideration, there are legal conditions preventing Statistics Netherlands from publishing individual data. The article by Al and Altena describes these legal preconditions and the problem of informed consent, and examines the security measures in detail.

Integrating administrative registers and household surveys

Paul van der Laan

1. Introduction

At Statistics Netherlands the design and organisation of the statistical process is undergoing a radical change. This change is prompted mainly by the growing demand for coherent statistical information, by rapid developments in information and communication technology, by high non-response rates for household sample surveys and by political pressure to cut down on staff costs and the reduce the reporting burden caused by statistics. The core of the new developments is a fundamental restructuring of the primary production process in which:

– *input* relies more on common sources (registers and multi-purpose data collected from a certain category of respondents) and less on data collected uniquely for a particular survey;
– *throughput* becomes a clearing house, making common input data useful for a wide range of statistical areas;
– *output* becomes well co-ordinated and available to a wide range of users through a corporate data warehouse function.

This article focuses on an important aspect of the throughput process: the integration of administrative registers and household sample surveys at the micro-level in order to create integrated micro-data files of persons, families, households, jobs, benefits and homes. In particular these micro-databases will be used to produce the data of the 2001 Dutch Census programme. As statistical information systems should provide accurate, relevant and authoritative information, the transformation of social statistics from a wide variety of largely isolated statistics into an integrated statistical system is the logical consequence of these prerequisites. The present article demonstrates why such a transformation will improve the quality of our statistical output. After all, it is in the common interest that the goals of government policy are carried out on the basis of the best available information.

The following section discusses the various approaches to producing social statistical information. Section 3 describes the traditional statistical process and discusses the drawbacks of the traditional approach. Next, section 4 presents the envisaged statistical production process and the place of integrated micro-data files in this redesigned process. Section 5 addresses the process of micro-integration: data editing and imputation. Section 6 discusses the sequence of micro-integration: the successive creation of micro-data files. Lastly, in section 7 conclusions are drawn.

2. Approaches to social statistical information

The increased demand from policy-makers for coherent statistical information is a major driving force behind the transformation of social statistics from a wide variety of isolated statistics into an integrated statistical information system. Integration does not have such a long history in social as in economic statistics, partly because there are no broadly accepted outcome indicators in the social domain. Nevertheless, various approaches to social statistical information have been developed during the past decades. [1)] These approaches can be summarised as:

a. data extracted from different isolated sources
b. data extracted from a comprehensive sample survey
c. data extracted from different isolated accounting systems
d. data extracted from linked accounting systems
e. data extracted from integrated micro-databases.

Each approach has its strong and weak aspects, as illustrated in Table 1. These will be further explained below.

a. Data extracted from different isolated sources

Traditionally, policy-relevant statistical data were mainly extracted from different isolated domain-specific sources, either administrative registers or household sample surveys. This method is simple, but has several disadvantages. Firstly, the accuracy of the sources is difficult to test and therefore the results are not undisputed. Secondly, it is not possible to estimate the relations between the data, as the information is derived from different sources. Thirdly, for the same reason, it is difficult to test whether the information on different policy indicators is consistent. Fourthly, it is hard to derive information on small population groups in society or on small areas, as the surveys are usually based on small sample sizes. Besides, as the information on groups in society was gathered from different statistical sources, definitions of those groups are rarely harmonised. Fifthly, also restricted by small sample sizes of the surveys, it is not possible to detect information on phenomena with a small incidence.

b. Data extracted from a comprehensive sample survey

Rather than being concerned with the distribution of aggregate variables over groups of statistical units, social research frequently focuses on relations between variables at the micro-level. This focus on the analysis of micro-data has been the driving force

Table 1
Strong(+) and weak(–) points of different approaches to social statistics

Aspect	Isolated sources	Compre-hensive survey	Isolated accounting systems	Linked accounting systems	Integrated micro-databases
Testing data accuracy (sampling errors and biases)	– –	–	+	+ +	+ +
Testing data consistency	– –	–	+	+ +	+ +
Making (multi-dimensional) cross-tabulations	– –	+ +	–	–	+
Information on small groups or small areas	–	–	+	+	+ +
Information on phenomena with a small incidence	–	–	+	+	+ +
Overall statistical coherence	– –	–	–	+	+ +

behind the development of ever more comprehensive household sample surveys covering as many variables as possible, so that all social relations can be analysed with a comprehensive micro-data set. This approach has its limitations. The response burden on the sampled households becomes too heavy if too many variables are covered in a single sample survey, although the change to computer assisted interviewing has facilitated the integration of different sample surveys into one combined electronic questionnaire. However, the tendency to retain existing indicators in a sample survey in order to make time series turns a sample survey into a sort of pressure-cooker: the response burden only increases. A more meaningful way to select indicators for a sample survey would be on the basis of their strategic use in policy and some kind of behavioural assumptions.

Although a comprehensive sample survey makes it possible to estimate the relationships between the indicators and the information derived from one sample survey is mostly consistent, a comprehensive sample survey still has serious disadvantages. Firstly, the accuracy of the comprehensive sample survey is, in most cases, not tested and therefore the results are open to question. Secondly, as with isolated sample surveys, it is hard to derive reliable information on small subgroups in society or on small areas, as the sub-samples unavoidably have limited sizes and suffer from selective non-response. Thirdly, also restricted by the small sub-samples, it is not possible to detect information on phenomena with a small incidence.

c. Data extracted from different isolated accounting systems

To produce a *coherent* data set in social statistics a statistical information system is needed in which the available information is combined in an integrating framework. Some countries have started to develop partial social accounting systems, for example in the form of labour accounting systems (Leunis and Altena 1996; Statistics Netherlands 1999; Walschots 1996).

Social accounting systems do not have many of the disadvantages of the analyses of separate data sources and large comprehensive sample surveys. Within accounting systems, the results of different data sources are confronted with each other and inconsistencies are eliminated. This improves the quality of the results and therefore makes them less arbitrary. However, as the sources are made consistent with each order on an aggregate level (by confronting tabulations), it still is not possible to produce reliable and valid estimations of the association between indicators from different sources at a micro-level. As within-cell joint variation of variables is disregarded, relationships between the variables in question need not be valid at a disaggregated level. Therefore, social accounting systems do not usually favour the analysis of micro-data. Furthermore, if the sources are surveys with common sample sizes, it is still not possible to derive information on small sub-groups, small areas or rare phenomena.
Apart from the 'within' disadvantages of these accounting systems, the most important disadvantage is that the integration work on several accounting systems will not automatically lead to consistent results between accounting systems. It is not unusual for different accounting systems to produce different results for the same policy-relevant indicator.

d. Data extracted from linked accounting systems

To cope with the problem of different social accounting systems producing different results for policy-relevant data, one suggestion would be to link accounting systems *on a macro-level.* As the national accounts produce authoritative statistical information, it could be promising to develop links between the System of National Accounts (SNA) and social statistics, thus facilitating the integration of social statistics by providing buoys. In practice, economic and social statisticians rarely work together to integrate the two fields of statistics. This does not mean that in theory there are no relevant links between the economic system and social statistics. As such, when partial social accounting systems like socio-demographic accounts, labour accounts, educational accounts, socio-economic accounts and social security accounts are consistent with one another and with the SNA, the resulting indicators will be mutually consistent as well. An example of such an approach is the development of Social Accounting Matrices, which extend the SNA with data on population, labour market, social security and other areas of social concern by integrating SNA data with data from available social accounting systems.

A second promising possibility to link social statistics to SNA lies within the common field of supply and demand of *services provided to persons* (Bakker and Van Rooijen, 2000). For example, social statistics on health describe the actual state of health of the population. However, the health status is not only influenced by demographic and socio-economic factors, but also by the consumption of health services. The production of health services is described by the economic statistics in volumes and values. In order to give a complete picture of health in society, the economic information on the production of health services has to be linked to the state of health, the related demand for health services and medical consumption.

Information on the quantity of the production of services (health services or others) can be obtained from business units or from persons and households. The production of services measured through a business survey should be similar to the consumption of services as measured in a household sample survey (for example the number of bed-days in hospitals). As such, we have two related integration processes. The first is the integration of the results on the quantity of the production of services from business and household sample surveys, and the second is the verification of the quantity of services by comparing it with the volumes calculated in the SNA. This is not a one-way process: the initial estimates of the volumes in the SNA can be used to compare with quantities measured and both sources can be adjusted (Bakker and Van Rooijen 2000).

What is there to be said about the advantages of linking information between accounting systems? The accuracy of all the sources used within both systems is tested and, therefore, the quality of the results is good. By integration of all the sources and the confrontation of the results of the accounting systems and consequently the evaluation of the production processes of these systems, the quality will improve further. Moreover the inconsistencies between the accounting systems are removed. However, there are still drawbacks to this approach.

First of all, compared with economic statistics, units in social statistics (persons and households) possess far more characteristics. This means that social accounting systems require more cross tabulations than the SNA, rendering an overall social accounting system very complex. In the second place, characteristics of units in economic statistics (such as economic activity, sector and size) are almost always included in (business) registers. On the other hand, many characteristics of units in social statistics have to be derived from sample surveys (e.g. educational attainment or occupation). Therefore many classification variables in social accounting systems are not available from the sample frame, i.e. they are not exogenous marginal totals in the statistical integration process. In the third place, as the sources are made consistent with each other on an aggregate level using techniques of statistical linking, it still is not possible to produce reliable and valid estimations of the relationships between indicators from different sources at a micro-level. Lastly, if the sources are surveys with common sample sizes, it is still not possible to derive information on small subgroups, small areas or phenomena with a small incidence.

e. Data extracted from integrated micro-databases

Developments in information and communication technology have created new opportunities for social statistics (Dunnet et al. 1999; Keller 1997; Keller et al. 1999), and increasing use can be made of data from large administrative registers. Furthermore, it is also possible to match data from administrative registers with data from household sample surveys. As different statistical data collections often show conflicting results for corresponding variables, matching different statistical data sources constitutes a powerful tool to enhance the accuracy of our statistics and avoid the publication of conflicting data [2]. There are basically three methods for matching micro-data from different sources.

The first method is *exact matching*: data on the same individual from different sources are linked. This requires identification of the individual by an identifier (matching-key) that is available in both sources. In the past decade Statistics Netherlands has acquired more and more experience with matching administrative records and household sample survey data. For example, since 1988 statistics on registered unemployment are compiled by linking register data of the employment agencies and survey data from the Labour Force Survey (LFS). The linked sample survey is subsequently weighted and raised using totals obtained from the register of the employment agencies (Kragt and Veenstra 1996). The foremost reason for using both sources is that the labour market status as registered in the files of the employment agencies is not very reliable. People are often still registered as being unemployed while they have already found a job. On the other hand, neither are the LFS data on the registration with an employment agency accurate enough to produce monthly figures. However, by combining the sources at a micro-level, more accurate information on registered unemployment is obtained on a monthly basis than could be obtained from both sources separately. Apart from using these large administrative data files for improving statistical information, the combined register information can also be used as a sample frame.

The second way of achieving linkage is *synthetic matching* or *statistical matching*. In this case common variables with a common breakdown are selected. The micro-data from all data sets are grouped according to this common breakdown. Within the resulting cells micro-data from different sources are combined randomly. This approach does not involve confidentiality problems and matching between samples is no problem either. The price to pay is that the validity of the micro-to-micro relationships derived from the synthetically matched data depends on the degree of association between the synthetic matching variables and the other variables. Within-cell joint variation is not picked-up. So, we are faced with the same drawbacks as in the case of data from social accounting systems.

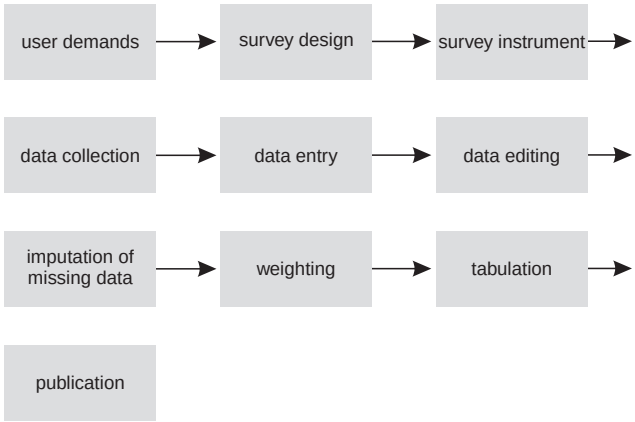
A third method of micro-to-micro linkage is to *redesign the sources* in such a way that there will be a joint questionnaire for a large sample that provides succinct information on core variables from all original surveys and more in-depth-questionnaires on the separate areas for smaller sub-samples. In this way a core micro-data set is obtained directly (the equivalent of exact matching). Combining the in-depth data from the various sub-samples by using the joint questionnaire as a synthetic matching-key can create an in-depth synthetic data set. Because the variables in the core are associated with the in-depth variables of the sub-samples, this approach picks-up relations more easily than synthetically matched data sets that rely on just demographic variables like sex and age for matching. Statistics Netherlands has developed an integrated system of social sample surveys using this modular approach: the *Quality of Life Survey* (Bakker and Winkels 1998; Winkels and Everaers 1998). The Quality of Life Survey questionnaire consists of a joint questionnaire for a large sample and subsequent questionnaires for sub-samples. The sample size of the large sample should be large enough to make it possible to give reliable information on small population subgroups, on small areas and on phenomena with a small incidence.

It is also possible to combine the techniques of record linkage and statistical accounting systems and build *integrated micro-databases* which will contain all the relevant information on persons, families, households, jobs, social security benefits and living quarters. This approach is central in this paper. These micro-databases are based on linked administrative and sample survey data and a reconciliation process called micro-integration. The methodology behind these integrated micro-data files is based on matching micro-data from administrative sources and household sample surveys and on a comprehensive use of techniques of statistical data editing, data imputation and sequential weighting. Central to this approach to statistics is the intensive use of *administrative data* (Al and Bakker 2000; Arts and Van Lith 1999; Blom and Carlsson 1999; Danmarks Statistik 1995; Van der Laan 1997; Vliegen and Van der Laan 1999). Once a statistical institute has access to administrative registers and can demonstrate that it uses them effectively, the statistical institute will also have better opportunities to gain influence over the administrative systems that it does not otherwise have, for example over their quality or perhaps even over their content. In many places in Statistics Netherlands' registers and household sample surveys are already combined in order to generate statistical data (See e.g. Arts 1996; Schaafsma-Harteveld 1999; Schulte Nordholt 1998*a*; Slagter 1999). Furthermore, by using register information as a weighting frame for sample survey data, sampling variance and non-response bias can be reduced. [3] Due to high non-response rates in the Netherlands, non-response has always been an important research topic. Suppressing non-response focuses on two aspects: reduction of non-response rates in the field and the treatment of sample survey data affected by non-response. Weighting is an important method to reduce the effects of non-response on sample survey data. Statistics Netherlands has developed new weighting methods where traditional weighting methods have failed. Other research on reducing non-response and on adjustment techniques is in progress (Geuzinge et al. 2000). Lastly, integrated micro-data files offer good possibilities to use synthetic and combined estimators for small subgroups and small areas.

3. The traditional statistical process

The traditional production process in social statistics is based on '*stovepipes*': a sample survey is designed with a questionnaire aimed at collecting the information needed to compile a specific publication serving specific user demands. The response to the questionnaires by persons or households in the sample is edited and subsequently entered in some kind of statistical package.

Figure 1
The traditional production process of social statistics



Publication totals are estimated by calculating one set of weights for the responding units. Every field of statistics has its own production process. Data collected for other fields are seldom used. The phases of the traditional production process are rendered in Figure 1.

This stovepipe approach has several weaknesses. First, it does not lead to coherent statistical output, as results on each subject are published independently of other subjects. Consequently, for a number of variables different outcomes are published by Statistics Netherlands. A certain variable may have different definitions in different publications. 'Labour cost' for example may be defined differently in trade statistics than in statistics on the manufacturing industry, as a consequence of which the two figures cannot be added up. On the other hand, labour cost data in the manufacturing industry may be unharmonised with employment data in the same industry, if business and employment surveys do not use of the same populations. The consequence in this case is that the figures on labour cost and employment cannot be related. Harmonisation of statistical units and concepts is a necessary but not a sufficient condition to solve this problem. Besides sampling errors, measurement and non-response errors also cause different outcomes. The most eye-catching lack of harmonisation is when different estimates are published for the same concept in different statistical publications. For instance, 'number of employees' is asked in almost any business survey and, hence, many estimates are obtained for the total number of employees for a certain branch of industry. Without further action all these estimates are likely to be different. For several reasons this is an undesirable situation.

The second weakness is that ideally a statistical institute should ask an enterprise only once for example about its labour cost and a household only once on its labour market participation; this is not the case in stovepipe approach. Furthermore, a lot of information is already available in the form of administrative registers for both enterprises and persons. At the same time, the mere existence of administrative registers bring about two major external incentives for a more efficient approach: political pressures to cut down staff costs and to lower the administrative burden caused by statistics. [4)] A third drawback is that the data editing and imputation processes are not performed optimally. Because only data are used that are collected in the sample survey in question, it is more time consuming to decide whether the data are accurate than when all the relevant data can be used that are available at the statistical institute.

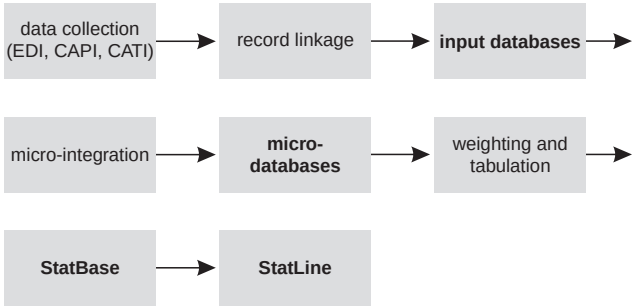
The last reason why the stovepipe approach is less attractive is the high non-response rate in household sample surveys. The average response rate in the Netherlands is around 55%. High rates of non-response invalidate the traditional weighting procedures, unless all assumptions about (conditional) unselectiveness of the non-respondents are valid, which is usually not the case.

4. Statistical process redesign

The production process of social statistics is changing constantly. Information and communication technology have considerably improved the data collection phase during the past decade (Wegman 1999). Techniques such as computer assisted personal interviewing (CAPI) and computer assisted telephone interviewing (CATI) have replaced paper questionnaires by electronic ones using laptop computers. Statistics Netherlands has developed special software for that purpose like Blaise® (Denteneer et al. 1994). Moreover, CATI and CAPI made it possible to integrate several tasks in the survey process, in particular data collection, data entry and data editing (Bethlehem 1996; Keller et al. 1999).

As pointed out in the previous section, the main disadvantage of the stovepipe approach is that it does not lead to coherent statistical output. In order to achieve coherence in our statistical output we need to redesign our statistical production process too. This implies replacing the traditional stovepipe production process by a production process where input will rely more on common sources and less on data collected uniquely for a particular statistical subject. [5)] The basic assumption is that the best data source can be chosen for every variable. Some variables are better observed in household sample surveys, others are better captured from administrative registers. [6)] The analysis of statistical data has to become a clearing-house, making common input data useful for a wide range of statistics. Apart from the phases where it is decided which data should be collected, the redesigned production process will consist of the phases shown in Figure 2. *Micro-integration* – the editing of individual records from different sources with the purpose of eliminating invalid, inconsistent or missing data – plays a central role in this process.

Figure 2
The redesigned production process of social statistics



The input databases (called Baselines) consist of all data from administrative registers on persons, families, households, jobs, benefits and living quarters. All these data are exactly matched. Baselines also contain all data from household sample surveys. All survey data are exactly matched with the administrative data sources. The micro-databases consist of mutually linked files on persons, families, households, jobs, benefits and living quarters. On behalf of our output programme, the data in the micro-databases are edited and made mutually consistent.

However, the micro-databases cannot directly be used as output databases. First of all, micro-data do not fulfil our confidentiality requirements. Furthermore, imputations might result in nonsensical data at the micro-level. To avoid these problems, the data must be aggregated to a suitable higher level. However, there is still a risk of unreliable data at certain low-level aggregates. All these considerations prohibit the direct and free use of these micro-databases as a data source for all kinds of statistical publications and analyses. Therefore, an additional database is needed, containing all publishable aggregates (over statistical units) of the micro-databases. This database is called StatBase. StatBase contains all statistical data of which Statistics Netherlands considers publication meaningful and reliable. Alongside statistical data, StatBase also contains meta-data. [7)]

Lastly, from StatBase is used to fill the output data warehouse: StatLine, a large number of multi-dimensional datacubes, each covering a theme (area of interest) and together providing a comprehensive and coherent picture of Dutch society. As themes may overlap, the same data appear in StatLine in different cubes under different themes. StatLine can be characterised as consisting of a set of standard views on StatBase. So, coherent statistical output denotes a situation where StatLine provides a comprehensive and coherent picture of society. This implies that there is consistency *within* datacubes and *between* datacubes.

Consequently, the focal points regarding the redesign of our statistical production process can be summarised as:
– Reduce the response burden in the data collection phase by more intensive use of data from administrative registers and primary and secondary electronic data interchange (EDI);
– Improve the accuracy of the statistics by improved data editing procedures (like graphical macro-editing and output editing) and non-response correction techniques;
– Reduce costs by using more cheap register data and fewer expensive sample survey data; [8)]
– Intensify the creation of integrated statistical data in the throughput phase by a combined use of linked register data and sample survey data and through application of new techniques of statistical data imputation and sequential weighting; [9)]
– Improve the access to statistical information by innovating the statistical output database StatLine so that thematic searchers are easily guided to the appropriate datacube.

The redesign calls for methodological investments in the field of social statistics aimed at the development of integrated micro-databases of persons, families, households, jobs, benefits and housing. The goals of these investments are:
– to ensure consistency between estimates from different sources;
– reduction of the effects of non-response;
– reduction of the sampling variance.

5. The process of micro-integration

The most important data sources at Statistics Netherlands containing micro-data on socio-demographic, socio-economic and socio-cultural characteristics of the population can be divided into three types: [10)]
a. *Administrative registers*: municipal basic administration of personal data (population register); administration of employee insurance schemes; public sector employees disablement benefits administration; social assistance benefits administration; employment agency files; students at vocational colleges and universities; income information system; address register of institutional households; housing register; valuation of real estate registration system; geographic base file.
b. *Business sample surveys*: annual survey on employment and earnings.
c. *Household sample surveys*: labour force survey; budget survey; socio-economic panel survey; continuous quality of life survey.
Each of the data sources mentioned above provides information on relevant aspects of the situation of the Dutch population. However, not all aspects are covered equally well. Furthermore, most administrative data sources suffer from biases and household surveys suffer from sampling and non-sampling errors. Especially in household sample surveys selective non-response and under-reporting are a notorious cause for inaccurate statistical results. In order to use the data sources efficiently in our new statistical production process they have to be linked, subject to the requirements that the number of matched records should be maximised and the number of mismatched records minimised. The matching process is described in Arts et al. 2000. [11)]

We do not yet use the income information system, the budget survey and the socio-economic panel survey in the present stage of the project on integrated micro-data files. The income information system consists of income and wealth data of persons and households obtained from various wage and income tax files (including the student grants administration and the individual rent subsidies administration). There are two reasons for this:
– Priority is given to the data needed to produce the 2001 census programme, i.e. no emphasis on income and expenditure data;
– Linking household sample survey data with tax data has not yet been authorised by the tax department of the Ministry of Finance.

After the creation of the input databases (Baselines) with all administrative and survey data on persons, families, households, jobs, benefits and housing exactly matched, we can start the reconciliation process by data editing and imputation. As we are faced with the problem that the data collected from registers or by sample surveys will always contain errors consisting of invalid, inconsistent or missing data, the main task in this phase is to check to what extent the linked data from the different sources are consistent with one another. The process of micro-integration consists of the following steps:
a. *harmonisation of units*: are the statistical units defined uniformly in all sources? (special reference to comparability in space and time);
b. *harmonisation of reference periods*: do all data refer to the same period or the same point in time?
c. *completion of populations (coverage)*: do all sources cover the same target population?
d. *harmonisation of variables*: are corresponding variables defined in the same way? (special reference to comparability in space and time);
e. *harmonisation of classifications*: are corresponding variables classified in the same way? (special reference to comparability in space and time);
f. *adjusting for measurement errors (accuracy)*: after harmonising definitions, do the corresponding variables have the same value?
g. *adjusting for missing data (item non-response)*: do all the variables possess a value?
h. *derivation of variables*: are all variables derived using the combined information from different sources? [12)]
i. *checking overall consistency*: do the data meet the requirements imposed by identity relations?

The aim of micro-integration is to check the matched data and modify incorrect records, i.e. data on individual units, in such a way that statistical analyses and publications based on the data in StatBase give acceptable results. Because the published results are aggregates, such as totals or means, it is not necessary to remove all errors. Small errors often cancel out when they are aggregated. Moreover, when the variables are obtained by means of a sample of the population, there will always be a small error in the results when they are extrapolated to the whole population, even if all collected variables are completely correct. This implies that for data obtained by means of a sample, an error in the results due to ‘noise’ in the data is acceptable if it is negligible in comparison with the sampling error.

To detect errors edit checks are specified by subject-matter specialists. A record that fails an edit is considered faulty, a record that satisfies all the edit checks is considered correct. The values of a faulty record should be adapted in such a way that the errors in these values are reduced or, preferably, even entirely removed. The combined use of administrative registers offers many possibilities to adjust for incorrect data. Some variables are more reliable in one register, others in another. In particular, missing units or missing values in one register can be completed from another register (sometimes by applying ‘marginal’ imputations). In this way a full coverage of the population in terms of micro-data is maintained. Another example of the advantage of using combined data is that, after matching with files of the administration of employee insurance schemes, some records in our annual survey on employment and earnings that were initially considered to be jobs, in fact turn out to be social security benefits paid through the former employer. Consequently, statistical data editing using the combined data of all the sources that need to be checked leads to a more efficient process of data editing than a stovepipe approach, where every data source is modified independently from the others. It takes less time to decide whether the records of a particular data source are accurate if all the relevant data can be used that are available at the statistical institute.

In addition to the modification of incorrect records, an important part in the process of micro-integration is the derivation of new variables based on the *combined* information from different sources (e.g. household status of a person; cf. step *h* above). These 'deterministic imputations' often yield relevant variables in the integrated micro-databases which cannot be derived – or not with equal reliability – in each data source separately.

Traditionally, statistical data editing is a very costly process, especially when a large number of records are involved as in the case of administrative registers. Statistics Netherlands has therefore devoted much research to automating the editing process (Pergamentsev 1998; Van de Pol and Bethlehem 1997; De Waal 1998). In order to produce reliable tables it is usually efficient to correct only the most influential errors. Therefore, it is interesting to test at several stages of the editing process what the resulting key tables would be if the editing process were stopped at those stages. By comparing these key tables at several stages it is usually found that the effect of editing quickly diminishes. As soon as major changes are no longer detected in the key tables, the editing process may be stopped. To increase the efficiency in statistical data editing Statistics Netherlands has developed the statistical software package CherryPi® (De Waal 1998), a general system for automatic editing and imputing data. An improved version of CherryPi® and other edit and imputation methods such as graphical macro-editing and the hot-deck imputation method are currently being implemented in a new software package called SLICE® (Statistical Localisation, Imputation and Correction of Errors). SLICE® is to be totally integrated in Statistics Netherlands' Blaise® suite (De Waal and Wings 1999).

How much micro-integration will improve the accuracy of our statistical output depends among other things on our *knowledge of invalid, inconsistent or missing data* and our *experiences with statistical techniques* (Holt 1999). Lack of relevant meta-data or insufficient experiences with particular statistical techniques can lead to the introduction of biases when adjusting for invalid, inconsistent or missing data.

6. The sequence of micro-integration

Not all files that are part of the whole set of integrated micro-data files can be created at the same moment in time. In order to manage the process we have to choose the order in which these files are created. Based on our present experiences, it seems useful to distinguish the following steps:

1. Create a temporary persons file based on administrative data (population register).
2. Create a temporary jobs file based on administrative data (annual survey on employment and earnings; administration of employee insurance schemes).
3. Create a temporary benefits file based on administrative data (administration of employee insurance schemes; public sector employees disablement benefits administration; social assistance benefits administration).
4. Create an address file based on administrative data (population register; housing register; geographic base file).
5. Create a temporary housing file based on administrative data (housing register; valuation of real estate registration system).
6. Match the temporary persons file with the address file.
7. Match the temporary persons file with the address register of institutional households.
8. Match the temporary persons file with the temporary jobs and benefits files.
9. Match the temporary persons file with the employment agency files.
10. Match the temporary persons file with the students at vocational colleges and universities file.
11. Match the temporary persons file with the household sample surveys (labour force survey; continuous quality of life survey).
12. Create a definite persons file. The number of persons in the definite persons file is normally equal to the number in the temporary persons file. Some characteristics of persons, however, may be adjusted. Lastly, all remaining missing basic characteristics of persons are imputed (sex, age, marital status, country of citizenship, country of birth, family status and household status).[13)] Other characteristics will be weighted.
13. Create a definite families file.
14. Create a definite private and institutional households file.
15. Create definite jobs and benefits files. Comparing jobs and benefits per individual person usually leads to dropping some jobs or benefits. The start and ending dates of jobs and benefits frequently suffer from administrative delay.[14)] Some categories of jobs are added from the labour force survey as they are not observed in administrations (off-record jobs like domestic helps, part-time medical domestics, clerics and religious community members, paperboys and outworkers). All remaining missing characteristics of jobs and benefits are imputed.[15)]
16. Create a definite housing file. All remaining missing basic characteristics of housing are imputed (type of housing and occupancy status).[16)] Other characteristics will be weighted.

The first five steps can be taken simultaneously and independently of each other. Once the temporary files are linked we can create the definite files using information from all matched files. Special attention has to be paid to records that cannot be linked, but have to be included in the definite files. All variables in the micro-databases that cannot be reliably imputed – especially those that stem solely from sample surveys – will be estimated in the weighting and tabulation phase. After the process of micro-integration, some inconsistencies in the micro-databases will still remain. This is not a real problem if the variables involved are not to be published. These remaining flaws will be adjusted during the weighting and tabulation process (See Kroese et al. 2000).
The reliability of the data in the micro-databases is determined by the following three factors:

1. The reliability of the sources used;
2. The reliability of the executed record linkages;
3. The reliability of the statistical techniques used.

The first exercise of the project on integrated micro-data files created a set of files referring to the situation on 31 December 1995 (Arts and Van Lith 1999). This exercise produced a persons file, a jobs file, a benefits file and a households file. The persons file contained 15.7 million records with 17 socio-demographic, 10 socio-economic and 2 socio-cultural variables. The jobs file contained 7.3 million jobs with 15 job characteristics. The benefits file contained 1.8 million benefits with 3 benefit characteristics. Every job and every benefit was related to a person in the persons file. The household file contained 6.3 million households.[17)] All persons in the persons file corresponded with a household in the households file. The paper by Arts and Van Lith (1999) presented some new statistical information based on the 1995 micro-data files. For example, data on jobs and earnings of employees by ethnicity and economic activity in the Netherlands were published for the first time.

For some variables the micro-databases contain information relating to the whole population (for example, sex, age, marital status, country of birth, family status, main source of income). These (independent) variables stem from registers. For some variables we only have information from sample surveys (dependent or target variables, for example educational attainment and occupation). In the 1995 exercise, methods of *mass imputation* of target variables were applied by using register variables as auxiliary variables. In step 12 above, not only the basic characteristics of persons were imputed but all characteristics. Variables not relating to the whole population were predicted using register variables as auxiliary variables.[18)] This approach led to a consistent set of estimates. The reason for using imputation instead of weighting was that we felt that

imputation was more flexible than weighting. When using weights, all the variables from a sample survey are inflated in the same way. In our experience of statistical integration some variables need to be inflated in one way, while others should be inflated in other ways (for example, educational attainment of a person observed in a sample survey is less biased than his or her income).[19] However, the applied imputation methodology has some serious drawbacks. Reliability is guaranteed only for a small (not comprehensive) set of such estimates. The method of imputation determines which estimates of dependent variables are reliable and which are not. After the imputations are made no flexibility is possible.

On the other hand, the more common approach of splitting the micro-database into various rectangular data sets and calculating *one set of weights* for each data set by means of the calibration method leads to reliable estimates. Consistency of the set of all estimates is not guaranteed, however, as there are usually not enough objects in the sample to satisfy all calibration equations needed for consistency. There are simply too many dimensions. Which estimates are consistent and which are not is completely determined after the weights are constructed. Analogous to methods of mass imputation, after the sets of weights are constructed no flexibility is possible.

As a consequence, imputation and weighting as described above are not valid methods to proceed from the micro-databases to StatBase. A more flexible approach has to be developed that combines the advantages of both methods. At Statistics Netherlands we are currently investigating methods of *sequential weighting*. In principle we are willing to adapt the weighting scheme for each estimation problem (say a particular table). If a new table is required (i.e. a table that is not yet in StatBase), the estimation is done as well as possible under the restriction that the resulting table has to be consistent with all tables already in StatBase. Each estimation problem is tackled by means of (re)weighting the relevant data set. Central in the sequential weighting approach are so-called minimal weighting models. First a set of starting weights is calculated for the sample survey (e.g. the Horvitz-Thompson estimator) and subsequently these weights are calibrated with respect to a set of register variables. The calibration restrictions are chosen such that consistency is obtained with the most important register variables (key variables), selective non-response is corrected for as well as possible and variances are minimised. Kroese et al. (2000) present the weighting and tabulation process leading to StatBase in more detail. Their paper also contains some results of the 1995 exercise.

7. Conclusions

As statistical information systems should provide accurate, relevant and authoritative information, the transformation of social statistics from a wide variety of largely isolated statistics into an integrated statistical system is the logical consequence of these prerequisites. It is in the common interest that social policy is carried out on the basis of the best available information. The creation of micro-data files of persons, families, households, jobs, benefits and housing based on a statistical integration of administrative registers and household sample surveys contributes to this objective. The production of integrated micro-data files as part of the redesigned production process of social statistics will lead to a number of benefits over the traditional stovepipe approach:

– our statistical outcomes will show more consistency;
– we have better tools to identify quality problems in our sources;
– we can reduce costs by using more cheap register data and fewer expensive sample survey data;
– we have better opportunities to produce data on small population groups;
– we have better opportunities for small area estimation;
– we have better opportunities to correct for the selectivity of non-response in household sample surveys.

An efficient social statistical system should contain both data from household sample surveys and from administrative registers. Statistics based on data from administrative registers are not a cheap, low-quality alternative to sample survey statistics; they have their own merits and are in many cases the best alternative. Administrative registers are used as a sampling base for household surveys, to complement sample survey data with register information, to produce information on small subgroups, small areas or phenomena with a small incidence and to analyse non-response in sample surveys. However, in order to use administrative data more effectively in our statistical system we must:

– encourage legislation that clearly enables statisticians to access administrative data;
– promote a better documentation, in particular the meta-data on administrative registers;
– invest in methodology of statistics based on administrative data or a combination of sample surveys and registers at the micro-level;
– regularly audit the quality of administrative data, especially in terms of coverage and administrative delay;
– try to gain influence over actions leading to modifications to an administrative file used for statistical purposes.

Notes

1) See for a review e.g. Bakker and Winkels 1998; Van Bochove and Everaers 1996; Kooiman and Van de Stadt 1991; Van Tuinen 1995; Van Tuinen et al. 1994.
2) Needless to say that such an approach makes high demands on privacy protection (Al and Altena 2000; Kooiman et al. 1999).
3) Register information is for example very useful to identify homogeneous population groups for applying techniques of post-stratification and calibration.
4) The reduction of the administrative burden on enterprises and institutions is stipulated in the Law on the Central Bureau of Statistics and the Central Commission for Statistics (Official Statistics Act 1996, Article 10).
5) This is particularly relevant in the case of administrative data sources. Statistics production can never be the primary concern when establishing an administrative record system. There is, therefore, almost always a need to combine information from various sources, before it can be published as official statistics.
6) Obviously, some variables can only be observed in household surveys.
7) The meta-data in StatBase will consist of *output meta-data* and *process meta-data*. Output meta-data specifies the statistical output (statistical units, classifications, concepts etc.). Process meta-data describes how the data are produced (sources and methods).
8) See also United Nations 1996.
9) The intensified use of administrative data also poses new challenges to statistical methods formerly centred on sample survey methodology.
10) The data sources containing micro-data on socio-demographic, socio-economic and socio-cultural characteristics of the population are briefly presented in the Appendix included in the end of this issue.
11) Matching on a personal identification number is very successful: 96-98% of the records can be matched. If such a identification number is absent, we match on address, date of birth and sex. This results in 93-95% of matched records.
12) Some of these variables will in fact be new, i.e. they can only be created, because we can use the *combined* data from different sources.
13) These imputation are of a negligible size as these characteristics stem from registers and registers as a whole contain hardly any missing values.
14) Administrative delay is the span of time between the occurrence and the registration of an event. See for a discussion of its statistical implications Hoffmann 1995.

15) See for some of the methods used Schulte Nordholt 1998*b*.
16) These imputation are usually negligible.
17) Following the international recommendations for the population and housing censuses in 2001 private households refer to the *household dwelling concept*, i.e. a private household is defined as the aggregate number of persons occupying a housing unit
18) As a matter of course, those register variables were chosen as auxiliary variables which had a high correlation with the variables to be imputed. Examples of those register variables are sex, age, ethnicity, household status, region, branch of economic activity, status in employment, time usually worked and type of benefit.
19) The means that the correlation between income and educational attainment in household surveys is also biased. This bias cannot be solved by putting weights to the sampled persons or households.

References

Al, P.G. and B.F.M. Bakker. 2000. 'Re-engeneering Social Statistics by Micro-Integration of Different Sources, An Introduction'. *Netherlands Official Statistics*, this issue.

Al, P.G. and J.W. Altena. 2000. 'Data-Security, Privacy and the SSB'. *Netherlands Official Statistics*, this issue.

Arts, C.H. 1996. 'Integration at the Micro Level of Data on Persons receiving National Assistance Benefits'. *Statistical Journal of the United Nations Economic Commission for Europe*, Vol. 13 (4), pp. 385–396.

Arts, C.H. and F.J. van Lith. 1999. 'A New Approach on Statistics about Persons: The Social Statistical Database (SSB)'. *Monthly Bulletin of Socio-Economic Statistics*, Vol. 16 (September 1999), pp. 22–29. *[in Dutch]*.

Arts, C.H. B.F.M. Bakker and F.J. van Lith. 2000. 'Matching Administrative Registers and Household Surveys'. *Netherlands Official Statistics*, this issue.

Bakker, B.F.M. and J. van Rooijen. 2000. 'One Number for the Supply and Demand of Services'. *Netherlands Official Statistics*, this issue.

Bakker, B.F.M. and J.W. Winkels. 1998. 'Why Integration of Household Surveys? – Why POLS?' *Netherlands Official Statistics*, Vol. 13 (Summer 1998), Special Issue, *Integration of Household Surveys: Design, Advantages and Methods*, ed. B.F.M. Bakker and J.W. Winkels, pp. 5–7.

Bethlehem, J.G. 1996. 'Efficient Survey Processing on Microcomputers'. In *Proceedings of the Expert Group Meeting on Innovative Techniques for Population Censuses and Large-Scale Demographic Surveys, The Hague, Netherlands, 22-26 April 1996*, Netherlands Interdisciplinary Demographic Institute and United Nations Population Fund, pp. 149–157. The Hague: NIDI; New York, NY: UNFPA.

Blom, E. and F. Carlsson. 1999. 'Integration of Administrative Registers in a Statistical System: A Swedish Perspective'. *Statistical Journal of the United Nations Economic Commission for Europe*, Vol. 16 (2/3), pp. 181–196.

Bochove, C.A. van and P.C.J. Everaers. 1996. 'Micro-macro and Micro-micro Linkage in Social Statistics'. In *The Future of European Social Statistics: Use of Administrative Registers and Dissemination Strategies; Proceedings of the Mondorf Seminar, Third session, Mondorf-les-Bains, Luxembourg, 25 and 26 January 1996*, Eurostat, pp. 205-212. Ed. Bernard Grais. Statistical Document. Theme 0: Miscellaneous, Series D: Studies and research. Luxembourg: Office for Official Publications of the European Communities.

Danmarks Statistik. 1995. *Statistics on Persons in Denmark: A Register-Based Statistical System*. Eurostat Statistical Document. Theme 0: Miscellaneous, Series D: Studies and analyses. Luxembourg: Office for Official Publications of the European Communities.

Denteneer, D., J.G. Bethlehem, A.J. Hundepool and M.S. Schuerhoff. 1994. 'BLAISE: A New Approach to Computer-Assisted Survey Processing'. In *Statistical Data Editing*, United Nations, Economic and Social Council, Statistical Commission and United Nations Economic Commission for Europe, Conference of European Statisticians. Vol. 1, *Methods and Techniques*, pp. 167–175. Statistical Standards and Studies No. 44. New York, NY: United Nations.

Dunnet, G., B. Halm and B. Sundgren. 1999. 'External Review of the Automation Function at Statistics Netherlands'. Voorburg and Heerlen: Statistics Netherlands. September 1999. *Mimeographed*.

Geuzinge, F.G., J. van Rooijen and B.F.M. Bakker. 2000. 'The Use of Administrative Registers to Reduce Non-response Bias in Household Surveys'. *Netherlands Official Statistics*, this issue.

Hoffmann, E. 1995. 'We Must Use Administrative Data for Official Statistics: But How Should We Use Them?' *Statistical Journal of the United Nations Economic Commission for Europe*, Vol. 12 (1), pp. 41–48.

Holt, T. 1999. 'Research and Development and the Future of Official Statistics'. *Research in Official Statistics*, Vol. 2 (1), pp. 21–31.

Keller, W.J. 1997. 'EDI, the Future'. *Netherlands Official Statistics*, Vol. 12 (Autumn 1997), Special Issue*, EDI: The State of the Dutch Art*, pp. 100–107.

Keller, W.J., J.G. Bethlehem, A.J. Willeboordse and W.F.H. Ypma. 1999. *Statistical Processing in the Next Millennium: The Impact of Information Technology on Data Collection, Processing and Dissemination*. Statistics Netherlands, Division for Research and Development, Research Paper No. 9917. Voorburg and Heerlen: Statistics Netherlands, Division for Research and Development.

Kooiman, P., J.R. Nobel and L.C.R.J. Willenborg. 1999. 'Statistical Data Protection at Statistics Netherlands'. *Netherlands Official Statistics*, Vol. 14 (Spring 1999), Special Issue, *Statistical Disclosure Control*, ed. J. Pannekoek and L.C.R.J. Willenborg, pp. 21–25.

Kooiman, P. and H. van de Stadt. 1991. 'Missing Data: Raising, Reweighting, Synthetic Estimation, Imputation, Synthetic Matching and Integration'. In *CBS Select 7: Statistical Essays: Statistical Integration*, Statistics Netherlands, pp. 119–135. The Hague: SDU Publishers.

Kragt, C.M. and C.J. Veenstra. 1996. 'Integration at the Micro-Level: Registered Unemployment'. Paper prepared for the Work Session on Registers and Administrative Records in Social and Demographic Statistics. United Nations, Economic and Social Council, Statistical Commission and Economic Commission for Europe, Conference of European Statisticians. Geneva, 11–13 November 1996.

Kroese, A.H., R.H. Renssen and M. Trijssenaar. 2000. 'Weighting or Imputation: Constructing a Consistent Set of Estimates Based on Data from Different Sources'. *Netherlands Official Statistics*, this issue.

Laan, P. van der. 1997. 'Census based on Integration of Administrative Data and Survey Data: The Dutch Experience'. In *Census Belgica 2001: Deux journées d'étude sur l'exploitation et l'avenir de recensement en Belgique, Bruxelles, 12 et 19 novembre 1996*, Point d'appui Travail, Emploi, Formation et Steunpunt Demografie. Actes volume 2, *Quel avenir pour le recensement en Belgique*, pp. 94–107. Louvain: Point d'appui Travail, Emploi, Formation.

Leunis, W.P. and J.W. Altena. 1996. 'Labour Accounts in the Netherlands, 1987-1993: How to Cope with Fragmented Macro Data in Official Statistics'. *International Statistical Review*, Vol. 64 (April 1996), pp. 1–22.

Official Statistics Act. 1996. *Law of 18 April 1996 establishing the Central Bureau of Statistics and the Central Commission for Statistics*. English translation published by Statistics Netherlands. Voorburg and Heerlen: Statistics Netherlands. April 1996.

Pergamentsev, S.Yu. 1998. *Automated Statistical Data Correction*. Statistics Netherlands, Division for Research and Development, Research Paper No. 9826. Voorburg and Heerlen: Statistics Netherlands, Division for Research and Development.

Pol, F.J.R. van de and J.G. Bethlehem. 1997. 'Data Editing Perspectives'. *Statistical Journal of the United Nations Economic Commission for Europe*, Vol. 14 (2), pp. 153–177.

Schaafsma-Harteveld, B. 1999. 'Disablement Benefits: Combining Survey Data with Register Data'. Supporting Paper prepared for the Joint ECE-Eurostat Work Session on Registers and Administrative Records in Social and Demographic Statistics. Working Paper No. 20. United Nations, Economic and Social Council, Statistical Commission and Economic Commission for Europe, Conference of European Statisticians, and Statistical Office of the European Communities. Geneva, Switzerland, 1–3 March 1999.

Schulte Nordholt, E. 1998*a*. 'Imputation, the Alternative for Surveying Earning Patterns'. *Netherlands Official Statistics*, Vol. 13 (Spring 1998), pp. 14–15.

Schulte Nordholt, E. 1998*b*. 'Imputation: Methods, Simulation Experiments and Practical Examples'. *International Statistical Review*, Vol. 66 (2), pp. 157–180.

Slagter, H.C.A. 1999. 'Compiling Structure of Earnings Statistics using Existing Survey Data and Register Data'. Invited Paper prepared for the Joint ECE-Eurostat Work Session on Registers and Administrative Records in Social and Demographic Statistics. Working Paper No. 7. United Nations, Economic and Social Council, Statistical Commission and Economic Commission for Europe, Conference of European Statisticians, and Statistical Office of the European Communities. Geneva, Switzerland, 1–3 March 1999.

Statistics Netherlands. 1999. *Labour Accounts in Theory and Practice: The Dutch Experience*. Voorburg and Heerlen: Statistics Netherlands.

Tuinen, H.K. van. 1995. 'Social Indicators, Social Surveys and Integration of Social Statistics: Strengths, Weaknesses and Future Developments of the Main Approaches in Social Statistics'. *Statistical Journal of the United Nations Economic Commission for Europe*, Vol. 12 (3/4), pp. 379–394.

Tuinen, H.K. van , J.W. Altena and H.C.M. Imbens. 1994. 'Surveys, Registers and Integration in Social Statistics'. *Statistical Journal of the United Nations Economic Commission for Europe*, Vol. 11 (4), pp. 321–356.

United Nations, Economic and Social Council, Statistical Commission, United Nations Economic Commission for Europe, Conference of European Statisticians and United Nations Economic Commission for Europe, Committee on Human Settlements. 1996. *Costing Aspects of Population and Housing Censuses in Selected Countries in the UN/ECE Region*. Statistical Standards and Studies No. 46. New York, NY: United Nations.

Vliegen, J.M. and P. van der Laan. 1999. Methodische und zeitliche Aspekte der Umstellung der amtlichen Statistik auf Register am Beispiel der Niederlande'. *Allgemeines Statistisches Archiv*, Vol. 83 (October-December 1999), pp. 434–446.

Waal, A.G. de. 1998. *An Introduction to CherryPi and Macro View*. Statistics Netherlands, Division for Research and Development, Research Paper No. 9836. Voorburg and Heerlen: Statistics Netherlands, Division for Research and Development.

Waal, A.G. de and J. Wings. 1999. *From CherryPi to SLICE*. Statistics Netherlands, Division for Research and Development, Research Paper No. 9908. Voorburg and Heerlen: Statistics Netherlands, Division for Research and Development.

Walschots, J.J. 1996. 'Statistical Information from Different Sources: Experiences with the Dutch Labour Accounts'. Paper prepared for the Eleventh Meeting of the Voorburg Group on Service Statistics. Newport, Gwent, United Kingdom, 16–20 September 1996.

Wegman, E.J. 1999. 'The Evolution of Statistics'. *Research in Official Statistics*, Vol. 2 (1), pp. 7–19.

Winkels, J.W. and P.C.J. Everaers. 1998. 'Design of an Integrated Survey in the Netherlands: The Case of POLS'. *Netherlands Official Statistics*, Vol. 13 (Summer 1998), Special Issue, *Integration of Household Surveys: Design, Advantages and Methods*, ed. B.F.M. Bakker and J.W. Winkels, pp. 8–11.

Linking administrative registers and household surveys

Koos Arts, Bart F.M. Bakker and Erik van Lith

1. Introduction

The first step in the statistical process of the SSB is matching the relevant sources, be they administrative registers or household sample surveys. As this matching will have to be successful to guarantee statistics of reasonable quality, the matching strategy should maximise the number of true matches and minimise the number of false ones. In this article we describe the development of a matching strategy for administrative registers on personal identification numbers and for household sample surveys on the combination of address, sex and date of birth.

Statistics Netherlands has some experience with linking administrative records and household sample survey data. For example, since 1988 statistics on registered unemployment are compiled by matching register data from the employment agencies and the labour force survey (LFS). The LFS is subsequently weighted and raised using totals obtained from the employment agency files. The reason for using both sources is that the agency files are too polluted to serve as a reliable source for registered unemployment, and that the LFS data on the registration with an employment agency are not accurate either. By combining the sources at a micro level, more accurate information on registered unemployment is obtained than could be obtained from both sources separately (Kragt and Veenstra, 1996).

The files of the population register form the backbone of the database, as all the other files are matched to this register. In this article, we describe the methods used to match the administrative registers and household sample surveys. Section 2 describes the sources and section 3 the strategy developed to match the data. In section 4 some theory for the estimation of the number of false matches is expounded and discussed. Section 5 contains the results of the matching process sand the conclusion is given in section 6.

2. The relevant sources

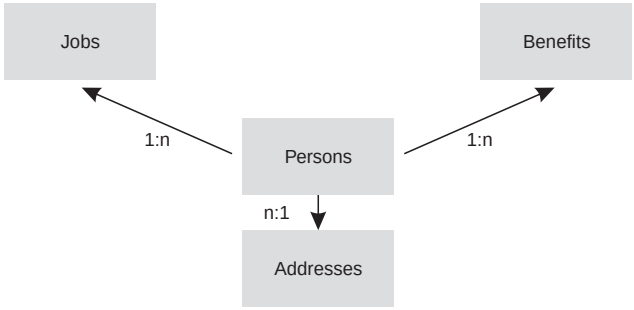
In order to produce relevant social statistics, information is needed on demography, jobs, social security benefits and dwellings. We therefore decided to match the most important sources summed up in the appendix included in the end of this issue for the year 1995:

- Population register (PR);
- Labour force survey (LFS);
- Continuous quality of life survey (CQLS);
- Health interview survey (HIS);
- Administration of employee insurance schemes, jobs (AEIS-jobs);
- Administration of employee insurance schemes, payroll data (AEIS-payrolls);
- Annual survey on employment and earnings (ASEE);
- Administration of employee insurance schemes, unemployment benefits (AEIS-UB);
- Administration of employee insurance schemes, disablement benefits (AEIS-DB);
- Administration of public sector employees disablement benefits (APEDB);
- Social assistance benefits administration (SABA);
- Register of addresses of institutional households (AIH);
- Key figures on census tracts and census districts (KFCTCD).

The quality of the data differs substantially. With the exception of the household sample surveys LFS, CQLS and HIS and the ASEE, all the sources are complete administrations. However, as the year 1995 was the first year that the registers were available for statistical purposes, some registers do not cover their target population because of non-response of some register owners. The quality of the available register data is sometimes also a problem.

We distinguish four measurement units: jobs, benefits, persons and addresses. Figure 1 shows that persons may have more than one job and may benefit from several social security payments and that more than one person can be resident at one address. The population register is the backbone of the SSB: all the other files are matched to it and in practice, the separate information on jobs, benefits and addresses are matched to each other via the population register. In our matching strategy, we choose to match the information on jobs and social security benefits with the aid of the information on persons in these jobs and receiving these benefits. Therefore, the matching strategy presented in section 3 focuses on the matching of records concerning persons. Matching of files containing addresses is also performed, but that strategy is outside the scope of this article.

Figure 1
Relations between jobs, benefits, persons and addresses



The target population of the SSB prototype is the total population of the Netherlands in 1995. As the population registers are dated on 1 January of each year, we need to combine the information from the registers of 1995 and 1996 to describe the target population of the year 1995.

- All the records of the 1996 population register are included, with the demographic and address information of 1 January 1996.
- Also included are the records in the 1995 population register of persons not included in the 1996 register, with their demographic and address information on 1 January 1995. These are the persons who died or emigrated in 1995.

The resulting file is called the ‘pooled population register’, which was created for the units ‘persons’ and ‘addresses’. By pooling the information, we almost cover the target population. The only people missing are illegal immigrants, some of the homeless and people who were born in or immigrated to the Netherlands and emigrated or died in the same year. For the efficiency of the matching procedure, it is relevant to note that some of the people who changed address in 1995 will not have the right address in the pooled population register.

In the Netherlands, the social-fiscal number (*sofi-number*) is used in many administrative registers to identify persons. However, in the 1995 and 1996 population registers this sofi-number is absent. It

was included for the first time in the 1997 population registers, so we matched the 1997 population register to the pooled 1995/1996 data. For the matching of the population registers, we used a unique personal identification number that is included in the population registers, the so-called 'A-number'. There is no sofi-number available for records of persons who emigrated or died between 1995 and 1997.

3. The matching strategy

We developed a matching strategy that maximises the number of matches and minimises the number of false matches. Beforehand, we decided that in order to have enough confidence in the outcomes of the matching process, more than 90 per cent of the records should be matched and not more than 5 per cent should be false matches.

The sources mentioned in section 2 contain all kinds of information which can be used in the matching process. The most important elements are:
1. sofi-number;
2. sex;
3. date of birth;
4. address (postal code, which has six characters in the Netherlands, and house number).

These variables can be used separately or in combination as 'matching keys'. In order to develop a matching strategy, the following aspects are relevant.
– The correctness of registration of the variable values that together form the matching keys. If there are errors in the registration of many values, relatively many false non-matches (non-matches which should actually be matches) and false matches (matches which should not be matches) will occur. We call this the 'quality' of the matching key.
– The extent to which the matching keys identify persons or addresses uniquely. When a matching key cannot identify persons or addresses uniquely, the number of false matches will be too large.

The following section will elaborate on the choice of the matching keys.

3.1 Matching on sofi-number

Matching on personal identification numbers has proven to be successful in Scandinavian countries (Longva and Thomson, 1996; Spieker, 1996). One condition for a high percentage of true matches is that the quality of the registered number is high. We assume that the quality of the sofi-number is high, as this number is used for administrative and fiscal purposes. Tax departments and employee insurance administration in particular require a high quality for their own tasks and duties. Moreover, the sofi-number is constructed in such a way that a check on the feasibility can be carried out. As many as 99.7 per cent of the available sofi-numbers in the pooled population registers fulfil this condition. As the sofi-number is a unique personal identification number, we expect the number of false matches to be very low. We prefer matching on the sofi-number to matching on the combination of sex, date of birth and address, assuming the latter will generate more false non-matches because of inaccuracies resulting from unreported changes of address and incorrect dates of birth for foreigners.

We are able to match on sofi-numbers for most of the administrative registers. In the annual survey on employment and earnings (ASEE), a combination of records collected through electronic data interchange and paper questionnaires, 16 per cent of records contain no information on employees' sofi-numbers. None of the household sample surveys contains the sofi-number. Together with the 16 per cent of the ASEE, these surveys have to be matched on the combination of sex, date of birth and address. The quality of the variables differs substantially. We tested several matching strategies empirically, in order to decide which is the most suitable.

3.2 Matching on other identifiers

Files that do not contain information on sofi-numbers are matched on combinations of sex, date of birth and address (postal code and house number), which we call 'other identifiers'. The aim is to find a matching key, or a set of matching keys to be applied successively, that leads to the highest possible matching rate, with the lowest possible rate of false matches. For the choice of the strategy, we distinguish the following matching possibilities:
1. Matching on the total set of identifiers (sex, date of birth, address) using the *present address.* Records are only matched if the values of all the 'other identifiers' are identical. The address information of the 1996 population register is used.
2. Matching on the total set of identifiers (sex, date of birth, address) using the *former address.* The records not matched in the first step are matched if the values of all the 'other identifiers' are identical when using the address information from the 1995 population register.
3. *Variations in sex and date of birth* are allowed. For records still unmatched after the first two steps, variations in sex and date of birth are allowed in cases of records with unknown date of birth (particularly foreigners) and records with erroneous registration of the date of birth. For each matching step, it is possible to distinguish between matching on the present and former addresses.
4. *Variations in address* are allowed. Lastly, variations are allowed on the postal codes and the house numbers. In these matching steps, it is again possible to distinguish between matching on the present and former addresses.

In each step, the first record with exact agreement on the particular subset of identifiers is matched, even if there are more possibilities (duplicate keys). To test which matching step leads to a high percentage of matches and to a relatively low percentage of false matches, we matched the data from the AEIS with those from the population register. The first contains approximately 5.5 million records of persons with a job or receiving a social benefit from the private sector. The latter contains 15.7 million records of Dutch inhabitants. Both the AEIS and the PR contain the sofi-number, of which we assume that it is of high quality and therefore will not lead to mismatches. Thus, we can consider records matched on the other identifiers to be a false match if the sofi-number of the AEIS differs from that of the PR.

The matching steps are performed consecutively: if a record is not matched in one step it becomes the input for the next.

We define a matching step as accurate if it increases the matching rate substantially and creates a low number of false matches. Therefore we excluded matching steps which did not increase the matching rate by more than 0.05 percent and steps with more than 10 per cent of false matches. Table 1 shows that only a few matching steps fulfil these conditions.

In the above described matching steps, many records of foreigners are not matched because of unknown dates of birth, entered as zeros in the population register. We tested whether steps 4g and 4h, in which the day and month of birth are variable, would lead to matches of reasonable quality when restricted to persons with u
We conclude that if no sofi-number available for matching, six matching steps are suitable for the matching of records from two different data-files on sex, date of birth and address. These are:
– matching on sex, date of birth and address on 1-1-1996;
– matching on sex, date of birth and address on 1-1-1995;
– matching on date of birth and address on 1-1-1996;

Table 1
Matching the AEIS with the pooled population register on sex, date of birth, postal code and house number, 1995 [1)]

Step	Kind of matching	Matched	Of which:	
			true matches	false matches
		%		
1	Identical on sex, date of birth and address 1-1-1996	82.7	99.7	0.3
2	Identical on sex, date of birth and address 1-1-1995	1.7	99	1
3a	Transposition of day and month of birth, otherwise identical (address 1-1-1996)	0.0	91	9
3b	Transposition of day and month of birth, otherwise identical (address 1-1-1995)	0.0	–	–
3c	Month of birth variable, otherwise identical (address 1-1-1996)	0.1	60	40
3d	Month of birth variable, otherwise identical (address 1-1-1995)	0.0	39	61
3e	Day of birth variable, otherwise identical (address 1-1-1996)	0.1	55	45
3f	Day of birth variable, otherwise identical (address 1-1-1995)	0.0	38	62
3g	Sex variable, otherwise identical (address 1-1-1996)	0.1	93	7
3h	Sex variable, otherwise identical (address 1-1-1995)	0.0	88	12
4a	Fifth character postal code variable, otherwise identical (address 1-1-1996)	0.0	91	9
4b	Fifth character postal code variable, otherwise identical (address 1-1-1995)	0.0	76	24
4c	Sixth character postal code variable, otherwise identical (address 1-1-1996)	0.1	98	2
4d	Sixth character postal code variable, otherwise identical (address 1-1-1995)	0.0	95	5
4e	House number variable, otherwise identical (address 1-1-1996)	0.5	98	2
4f	House number variable, otherwise identical (address 1-1-1995)	0.0	94	6
4g	Day and month of birth variable, otherwise identical (address 1-1-1996)	0.5	43	57
4h	Day and month of birth variable, otherwise identical (address 1-1-1995)	0.1	22	78
	Total	85.7 [2)]	99.2	0.8

1) The percentage of matched records is calculated as the percentage of the total number of records in the AEIS (5.5 million).
2) The relatively low percentage of matches (85.7%) can be explained by the fact that the register of the administration of employee insurance schemes for jobs and benefits for the year 1995 was delivered to Statistics Netherlands in 1997. Therefore addresses in the register have been updated up to 1997 resulting in many false non-matches. When matched on the sofi-number, more than 97 percent of the records of the AEIS is matched with the PR (see table 5), which argues in favour of the preference for the sofi-number stated in section 3.1.

– matching on sex, date of birth, first five characters of the postal code and house number (address on 1-1-1996);
– matching on sex, date of birth and postal code (address on 1-1-1996);
– matching on sex, year of birth and address on 1-1-1996, restricted to persons with unknown day and month of birth in the population register.

4. Estimation of false match rates

Section 2 described how the occurrence of false matches is tested by comparing the sofi-numbers in the two matched records in order to gain an insight into the quality of the matches made in certain matching steps. In general, this sofi-number is not present in the household survey file to be matched. Moreover, if it were, it would be preferred as a matching variable, not a test variable. So we would like to develop a method to estimate the expected number of false matches for a certain matching strategy, independent of the matching variables used, and without the use of test variables. In this section we discuss the two mechanisms that lead to false matches and try to develop an estimation method for the expected number of false matches resulting from it (sections 4.2 and 4.3). It will appear that this estimation cannot be made properly for one of the mechanisms. Alternatives are mentioned and suggestions for further investigations are made (sections 4.4 and 4.5).

4.1 Causes of false matches

The two mechanisms that may lead to false matches are:

– The actual presence of identical matching key values, e.g. same-sex twins living at the same address.
– Incorrectly registered matching keys in one of the matched files, that result in a matching key that is identical to a matching key of another unit.

In theory, the magnitude of these two classes of false matches can be estimated separately and added together to get the total. The estimations described below both apply for a single matching procedure, i.e. with the use of one matching key. If a matching strategy has several matching steps, as for instance is shown in Table 2, these estimations have to be made separately for each key. If the successive steps are applied to *the remaining records* in both files, the total expected number of false matches is the sum of all the estimations.

4.2 Identical matching key values

The number of false matches caused by the actual presence of identical matching keys is estimated with the aid of the number of identical keys. We assume that the matching process is carried out with files A and B, where the population of B (e.g. a sample survey) is a subset of A (e.g. the population register), the situation for almost all matching processes in the SSB. If we then have two records in file A and one in file B with identical keys, the matching of only one of the records in A results in a true match, the other match is false. Since we have no means to decide which match is correct, we simply choose one, with a 50% chance of making a false match. If there are three records in the population register with identical matching keys, the chance of a false match is 67%. If the two matched files both contain more than one identical matching keys, the calculation is somewhat more complex but analogous.

In general we can write down the calculation as following. Assume that we match file B to file A, the population of B being a subset of the population of A. Assume that a particular matching key occurs *m* times in file B and *n* times in file A, and that $m \leq n$ (if $m \geq n$, *m* and *n* can be changed). This will result in *m* matches. For each of these matches, the chance of a true match is *1/n* and the chance on a false match *(n-1)/n.* For *m* matches on this matching key, the expected number of false matches will be *m*(n-1)/n.* This leads to an expected total number of false matches of:

$$n_{fm,i} = \sum_{i=1}^{n_{lk,B}} m_{s_i} \frac{n_{s_i} - 1}{n_{s_i}} \qquad (1)$$

in which:

$n_{fm,i}$ total number of false matches because of identical matching keys,
$n_{lk,B}$ number of different matching keys in file B,
s_i the matching key of the i^{th} record in file B,
n_{si} the number of records in file A that match on key s_i,
m_{si} the number of records in file B that match on key s_i.

4.3 Incorrectly registered matching key values

The second cause of false matches – errors in the registration - leads to one of two situations: say there is an error in the registration of a matching key in file B, either the falsely registered key value is present in file A, or it is not. In both situations the error causes a false non-match, since the correct pair of records is not matched, but in the first situation it also results in a false match.

Assume again that we match file B to file A and that we consider the errors in the matching key in file B. More correctly we should say: we consider deviations of the matching keys compared with the values in the corresponding records in file A. But since our aim is to predict the expected number of false matches *in the matching of these two files* caused by errors in the registration, this is a correct approach.

We can discuss the estimation by dealing separately with the chance that an incorrect registration occurs and the chance that such error will cause a false match. This can be written down as follows:

$$n_{fm,e} = n_B \times P_e \times P_{ef} \qquad (2)$$

with:

$n_{fm,e}$ estimated number of false matches caused by errors in the registration;
n_B number of records in file B;
P_e chance that an error in the *matching* key occurs;
P_{ef} chance that an error causes a false match.

Since we have assumed that file B is a subset of file A the following reasoning applies. An optimal *matching* procedure would lead to 100% matched records. In that case every false non-match is caused by an error in the *matching* key. As we have seen, errors in the key also lead to false matches, but if the *matching* procedure is a proper one, i.e. the *matching* key is strongly identifying, the number of possible key values is much larger than the number of existing values in file A. Therefore the number of false matches caused by errors in the matching key is much smaller than the number of false non-matches. Moreover, part of this effect is accounted for by the estimation of the number of false matches due to identical *matching* key (section 3.2), i.e. those errors that not only result in values that exist in file A but also in file B.

So in this case we can estimate the chance of a false match by the percentage of false non-matches:

$$P_e \approx \frac{n_B - n_{B,m}}{n_B} \qquad (3)$$

with:

$n_{B,m}$ the number of matched records of file B.

The estimation of the chance that an error in the registration leads to a false match, P_{ef}, is more complicated and requires quite a detailed knowledge of how the register is handled by its day-to-day users and/or how they enter and update data (municipal staff for the PR, employees of the benefit agencies for the AEIS, of Statistics Netherlands interviewers for the sample surveys). This knowledge is necessary to make estimations for the correlation of certain mistakes, how a mistake changes a value and so on. One could think of features like:

– The chance of an error occurring is not spread randomly over the file, but correlates with certain quantities. For instance: people who move house often, often have their address changed and so the chance of an error occurring is larger than for someone who remains at the same address for years on end.
– The chance that an error will lead to a false match also correlates with certain quantities; for instance: consider a campus, where the postal codes of the living units resemble each other, and the where the years of birth of the residents (students) are concentrated in a small range. Here the chance that an error in 'month of birth' will lead to a value of another person (i.e. another student living on the campus) is much larger than on average in the population.
– The chance of an error in one variable is not independent of the chance of an error in another variable in the same record. An example is a foreigner who does not speak Dutch very well. The chance that his address is wrongly entered in the file is larger than for a native speaker. The chance that the date of birth of such a non-native is entered wrongly or is partly unknown is also larger. Both mistakes strongly correlate, since they have a common cause.
– The chance of a value for a certain variable changing to another one is not random, for instance:
  – controlling mechanisms in administrative computer software allow only existing postal codes;
  – keyboard logistics mean that a '2' will more often be mistakenly entered for an intended '1', than say a '9';
  – the effects of letters that sound similar when spelling something out. The best example is probably the exchange of a 'm' and a 'n' and vice versa in the postal code.

We have carried out tests to see what happens if one simply disregards these kinds of features and assumes that errors in the registration occur at random and independently of each other. This provides insight since in that situation it is easy to estimate P_{ef} by n_A/n_{poss} (n_A being the number of key values in file A, n_{poss} = the product of the numbers of all possible values for each variable in the matching key). The only useful result of the test is that the order of the expected number of false matches is the same as found with the use of the sofi-number as test variable (Table 2). The conclusion is that one has to account for the features mentioned above.

4.1 Incorrectly registered matching key values, alternatives

Another, more widely known strategy is to perform the matching of file B to A without using the variable investigated. This strategy can even be applied if B is not a subset of the population of A. In the matched file we compare the values of this variable as they were registered in the two files. The percentage of cases in which these values are not identical is a measure for the chance of an error in the variable leading to a false match. For example, if we perform the matching without using the variable 'sex' and find 90% identical values in the matched file, and similarly we find 95% identical values

for the variable ‘age’, the total expected chance of making a false match because of errors in ‘sex’ and ‘age’ is

1 – ((1 – 0,90) * (1 – 0,95)) = 0,995.

For the purpose we are discussing here, this method is not useful either, for two reasons:
– Firstly the assumption is, again, that mistakes in the key variables appear independently of each other, which is not the case, as explained above.
– Secondly the method assumes that when performing the matching step without the variable studied (‘sex’ and ‘age’ in the example) no false matches are made. But if they do occur and a different value for ‘age’ is found, the reason may very well be that two different persons are concerned. Counting this as an error in the variable ‘age’ overestimates the number of errors and thus the number of false matches because of errors in ‘age’. Table 2 shows that this problem shows up for several variables, for instance in the steps 3c, 3d (month of birth), 3e en 3f (day of birth) and 4a and 4b (5th character in the postal code).

We can conclude that methods other than those presented and tried here should be investigated. One could think of the use of frequency tables of variable values and their crossings. Another idea could be to investigate the correlation of errors in key variables, with the use of the sofi-number as test variable.

5. The results of the matching procedure

This section presents the results of the matching strategy developed in section 3: the construction of the pooled population register and the outcome of the matching of all the files mentioned in Table 1, except those describing the unit ‘address’.
Table 2 gives the results for the construction of the pooled population register. By pooling the 1995 and 1996 population registers, a file with 15.7 million records of persons is created. About 219 thousand records are included only in the 1995 population register (people who died or emigrated in 1995) and 294 thousand only in 1996 (those who were born or who immigrated in 1995).

The sofi-number of the 1997 population register was added to the pooled population register by matching the files with the use of the A-number. In 1995 and 1996 a number of records have the same A-number and are excluded from this procedure, because the chances on false matches are too high. Almost 98 per cent of the records receive a sofi-number. The matching rate differs between stayers, inflow and outflow. For the stayers the matching rate is 98 per cent and for the inflow 94 per cent. The matching rate for the outflow is rather low (6%). Only for those people who returned in 1996 after emigrating in 1995 could a sofi-number be added. This means that there will be some selectivity of the matches on sofi-number, as a very low percentage of records of persons who died or immigrated during 1995 or 1996 contain a sofi-number in the population register.

Table 3 describes the results for the matching between the other files and the pooled population register. Records of people who do not live in the Netherlands are excluded from the calculation of the matching rate. The matching rate for files containing the sofi-number is between 96 and 98 per cent. For the files that do not contain a sofi-number and are matched to the population register on other identifiers, the matching rate is between 93 and 95 per cent. Almost 82 per cent of the ASEE, of which 84 per cent of the records contain a sofi-number, matched on this personal identification number. The remaining records of this file were matched using the other identifiers. For a large part of the remaining records no information on the postal codes and house numbers was available and only 64 per cent could be matched. This brings the total matching rate to 93.4 per cent.

All in all, the overall matching rate is high enough to for us to have confidence in the quality of the results. Nevertheless, the confidence will be greater if the matched records are representative for the total population. Therefore, we analysed the selectivity of the matching process by comparing the frequencies of some of the key variables of the matched and non-matched records. The results are described in appendix 1. From the results we can conclude that:
– The matching rate for records of men is lower than that for women, in particular in administrative registers.
– The matching rate for records of young adults (15-24 years of age) in household sample surveys is lower than that for other age groups. This is also the case for the records of those who have never married. The reason for this is that people in these groups are more mobile and are therefore relatively often registered at a wrong address.
– In the SABA file, records of persons older than 65 and widowers have a lower matching rate than others.

Table 2
Results for the construction of the pooled population register, 1995

	N-records	1995 and 1996 = stayers	1995 only = outflow	1996 only = inflow	No unique A-number	Total
	x 1,000					
PR 1995	15,424	15,198	219	–	8	15,424
PR 1996	15,494	15,198	–	294	2	15,494
Pooled PR	15,710	15,198	219	294	0	15,710
Matching sofi-number From PR 1997						
Pooled PR	15,710	14,890	13	277	0	15,189
	%					
Matched records		98.0	5.7	94.2	–	96.7

Table 3
Results of the matching of other files to the pooled population register, 1995

	N-records	Population	Matched using			Matched Records
			Sofi-number	Other	Total	
	x 1,000					*%*
AEIS-jobs	8,358	8,336	8,114	[1]	8,114	97.3
AEIS-payroll	7,437	7,417	7,219	[1]	7,219	97.3
AEIS-DB (incl. DBG)	962	934	899	[1]	899	96.3
AEIS-UB	849	841	820	[1]	820	97.5
SABA	552	552	535	[1]	535	97.0
ASEE	1,662	1,653	1,350	193	1,543	93.4
LFS	191	191	–	179	179	93.8
CQLS	4	4	–	4	4	93.8
HIS	10	10	–	9	9	95.0

[1] No matching carried out.

- The matching rate for records of residents of the four large cities is lower than that of others.
- The matching rate for records of ethnic minorities is lower than that of others.
- The matching rate for records on benefits or jobs that started recently is lower than of others.
- In the ASEE, records of 'small' jobs are matched more often than those of others.
- In the CQLS, records of people who frequently visit a museum are matched more often than those of others.
- In the HIS, records of persons who were admitted to hospitals were matched more often than those of others.

6. Conclusion

In 1997 a research programme started to explore the possibilities of compiling statistics on persons and households by matching administrative registers and household sample surveys in order to provide the data for the population census of 2001. After matching the sources, inconsistencies in the data are examined and corrected by integration methods. The problem of missing data is solved by imputing and weighting. The present article focused on the matching of the sources. In conclusion, the proposed working method looks promising. Matching on a personal identification number shows to be successful: between 96 and 98 per cent of the records can be matched. If such an identification number was absent, we matched on sex, date of birth and address (postal code and house number). This results in a matching rate of between 93 and 95 per cent.

The sample frame of the 1995 sample household surveys consisted of addresses, from which interviewees were selected through different methods. This procedure leads to a high number of non-matches, namely for people who move house twice in a year and those not correctly registered in the population register. This was one of the reasons we decided to change our sample frame into the population register itself. The initial results of matching records sampled from the population register show a matching rate of almost one hundred per cent.

We tried to develop a method to estimate the number of false matches. The expected number of false matches caused by the actual presence of identical matching key values can be estimated correctly. The expected number of false matches caused by incorrectly registered matching keys in one of the matched files turns out to be more problematic. As a result the total expected number of false matches does not correspond to the number actually found, because the assumption is violated that incorrect registrations occur at random. Further research will be necessary to estimate the expected number of false matches more adequately.

References

Kragt, C.M. and C.J. Veenstra, 1996, *Integration at micro level: registered unemployment*, (Geneva: Paper Work Session on Registers and Administrative Records in Social and Demographic Statistics. Statistical Commission and Economic Commission for Europe).

Longva, S. and I. Thomson, 1996, *Reducing costs of censuses in Norway through use of administrative registers* (Geneva: Paper Work Session on Registers and Administrative Records in Social and Demographic Statistics. Statistical Commission and Economic Commission for Europe).

Spieker, F., 1996, *Output from a register-based statistical system*. (Geneva: Paper Work Session on Registers and Administrative Records in Social and Demographic Statistics. Statistical Commission and Economic Commission for Europe).

Appendix 1
Matching rate of the matching process by sex, age, ethnicity, region, marital status and some target variables, SSB 1995

	LFS	CQLS	HIS	AEIS-payroll	ASEE	AEIS-UB	AEIS-DB	SABA
	%							
Sex								
Man	93.7	93.9	94.7	97.3	92.5	97.0	95.8	96.5
Woman	93.9	93.6	95.2	98.1	94.3	98.1	97.2	97.5
Unknown	33.3	n/a	n/a	83.9	n/a	92.0	95.0	n/a
Age								
0–14	94.1	n/a	95.8	95.1	37.0	100.0	50.0	75.3
15–24	90.4	87.2	91.0	97.8	94.7	97.6	97.4	96.4
25–34	93.6	94.4	94.6	97.2	93.0	97.0	97.2	97.1
35–44	95.3	96.5	95.4	97.9	93.2	97.7	97.0	97.7
45–54	95.0	93.6	95.2	98.2	93.5	98.0	96.6	97.7
55–64	94.5	94.1	95.8	97.6	93.4	97.9	95.5	97.2
>65 yrs	94.2	93.0	96.6	94.1	93.1	96.0	93.3	88.9
Unknown	0.0	n/a	n/a	83.8	100.0	84.9	n/a	85.1
Ethnic group								
Native Dutch	94.4	92.7	95.2					97.3
Other	87.7	88.7	91.2					94.9
Unknown	n/a	94.7	n/a					100.0
Region								
Four large cities	90.4	92.8	93.6	97.8	82.6	96.8	95.3	96.8
Municipalities 100,000 inh.	92.8	92.8	93.1	98.5	91.4	97.4	96.5	97.1
Other	94.4	94.1	95.5	98.5	96.2	97.7	96.6	97.0
Unknown	n/a	n/a	n/a	80.6	n/a	0.0	5.8	n/a
Marital status								
Married	95.3	94.8	95.9					96.8
Divorced	93.3	93.8	94.1					97.8
Widowed	92.6	91.7	96.6					89.5
Never married	91.1	92.3	93.8					97.0
Unknown	94.0	100.0	n/a					n/a
Labour participation								
Active labour force	94.4							
Inactive labour force	94.0							
Other (incl. unknown)	93.3							
Museum visits								
More than once a month		95.9						
More than 3 times a year		94.6						
Less than 3 times a year		93.8						
Never		93.2						
Unknown		100.0						
Hospital admission								
None			94.9					
1 admission			95.4					
2 admissions			98.5					
3 admissions and more			100.0					
Unknown			90.6					
Days worked for social security								
0				93.8				
1– 99 days				96.6	96.4			
100–199 days				97.3	97.0			
200 days				98.3	92.6			
Starting date benefit								
Before 1995						97.7	96.3	97.7
In 1995						97.4	96.2	95.7

Weighting or imputation: constructing a consistent set of estimates based on data from different sources

Bert Kroese, Robbert H. Renssen and Marjolijn Trijssenaar

1. Introduction

The design and organisation of the statistical process is changing rapidly at Statistics Netherlands. This change is motivated by the need to produce more consistent data and to reduce the response burden, and by political pressure to cut down staff costs.

Keller et al. (1999) give an outline of the statistical process in the near future; this is presented briefly in section 2. A number of methodological problems have to be solved before the new statistical process can be implemented, particularly in the field of finding an appropriate estimation technique. One possibility to generate estimates is mass imputation, but generally speaking this technique may lead to unreliable estimates as will be explained in section 3.

In section 4 a weighting approach to estimation is proposed. Several rectangular micro-datasets are constructed, each of which is used to obtain a specific set of estimates by means of weighting. Here we allow the same rectangular micro-dataset to be weighted several times. It is argued that this strategy is appropriate for the new statistical process.

The present article will focus on person and household statistics. In section 5 some preliminary testing results are discussed of a test in which census-like tables are estimated using a prototype micro-database for persons.

2. The new statistical process

2.1 Drawbacks of the old production process

Traditionally, the production process of a statistical publication is designed along the lines of a 'stovepipe' model. A questionnaire is designed with questions aimed at collecting the information needed to compile a particular publication serving specific user needs. The questionnaire is sent to the units (mostly businesses, households or persons) in the sample and the response is edited and subsequently entered in some kind of statistical package. Publication totals are estimated by calculating one set of weights for the responding units.

In the past, there were many isolated stovepipes in our institute, making harmonisation very difficult. Without further action there is no guarantee that the estimates resulting from the various stovepipes can be related to each other. It might very well be the case that different estimates for the same concept are published in different statistical publications. The need for more conceptual and numerical consistency, and the increased availability of data from administrative registers, are the main reasons for a change in the statistical process.

The change was also motivated by cuts in the budget of Statistics Netherlands and the political desire to lower the response burden, in particular for businesses.

2.2 The new statistical process

This section gives a short description of the new statistical process as envisaged in our institute for the near future. A more detailed description can be found in Keller et al. (1999).

Figure 1
A diagram of the new statistical process

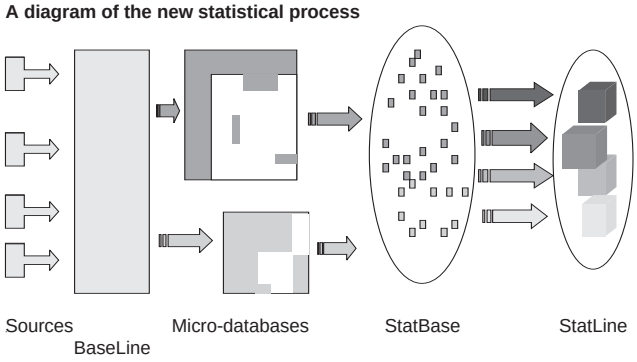


Figure 1 is a diagram of the new statistical process. Data from different sources (administrative registers, sample surveys, electronic data interchange (EDI)) are matched in the input-database '*BaseLine'*. The data in the sources refer to observation units which are not necessarily equal to statistical units. In BaseLine all input-data are adapted to data on statistical units.

In a second step, a *micro-database* is constructed for each object-type (persons, households, businesses etc.). For example, the micro-database for persons can be seen as a table with a row for each person in the population. Each column holds a variable. The data in the micro-databases are edited, partial non-response is imputed, and the variables are harmonised. The values of some variables are known for (nearly) all persons in the population (sex, age, marital status); the values of others are only known for a subset of the population, for example a survey sample. Consequently, the micro-database contains many empty cells.

The next step in the new process is estimating population totals based on the data in the various micro-databases. If the values of a variable are known for all persons, the total for the whole population is calculated by simply adding all values. If the values are known for only part of the population a more advanced statistical estimation technique has to be applied. All estimates are placed in *StatBase*, which can be regarded as a database containing all the data that Statistics Netherlands considers worthwhile publishing. The set of estimates in StatBase should satisfy the requirements of:

– reliability: the estimates should either be generated by means of an approximately design-unbiased method (variances reasonably small) or by a model-based method where the model used is plausible in some sense;
– consistency: no contradictions may result from confrontations of estimates in StatBase; for example, it should not be possible to derive two different 'total number of 15 year old children in a certain municipality' by combining estimates;
– disclosure control: no estimates can be derived that reveal individual information.

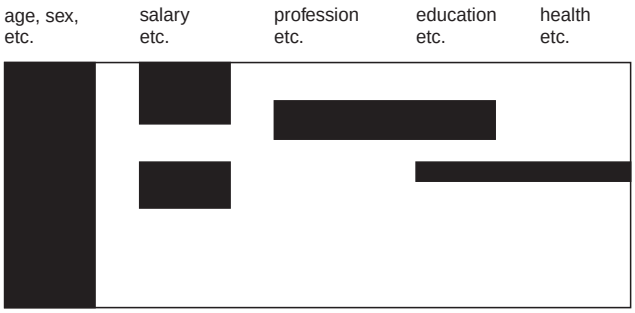
The estimates are disseminated by means of *StatLine*, which can be seen as a 'view' on StatBase. StatLine is user-oriented, i.e. the aggregates are arranged in multi-dimensional tables (data-cubes) that reflect standard 'areas of interest' or 'themes' in society. For example, it is our ultimate goal to gather all data referring to the theme 'health' in one such a data-cube, from which users can select the data they are interested in.

2.3 The estimation problem

A number of methodological problems have to be solved before the new statistical process can be put into practice. One of the main methodological challenges is how to proceed from micro-databases to StatBase. An estimation technique has to be developed that starts out from a number of only partly filled micro-databases and results in a set of reliable estimates that are consistent, satisfy the requirements of disclosure control, and are sufficiently comprehensive to satisfy user needs.

Part of this question is the core topic of this paper. Not much attention will be paid here to the confidentiality condition. In person and household statistics population tables of interest usually relate to a large set of persons and hence fulfil the disclosure control requirements. So we shall focus on the question of how to ensure that the conditions of reliability and consistency are met.

Figure 2
A simplified micro-database for persons; 'black' indicates data, 'white' empty cells



To be specific, we consider a simplified version of the micro-database for persons as presented in Figure 2.

The backbone of this micro-database is the population register (municipal basic administration of personal data) containing variables like 'age', 'sex', 'municipality' and 'marital status'. The persons contained in this register constitute the population of the micro-database and hence, by definition, the variables in the register are known for all persons in the micro-database. Matched with the population register is an (incomplete) register containing information about salaries etc. of a large number of people employed in the private sector. In the simplified micro-database presented in Figure 2, two sample surveys are also included. One sample survey (n=165,000) is the labour force survey and contains variables like 'occupation' and 'status of employment'. The other survey (n=10,000) contains variables related to health. Both surveys record the educational status of the respondents. No persons are included in both sample surveys. Some persons are included in one of the sample surveys and in the register with salary information.

Given a micro-database, we distinguish two estimation methods to fill StatBase. One is based on (mass) imputation and the other on weighting techniques. In principle, methods based on mass imputation generate mutually consistent estimates without guaranteeing reliability, whereas weighting methods generate reliable estimates without guaranteeing mutual consistency. In order to meet both the consistency and the reliability requirement, we suggest a new weighting strategy, one that can be seen as a new application of old techniques. It should lead to a sufficiently large set of mutually consistent and reliable estimates. Before we present this strategy the method of mass imputation is discussed.

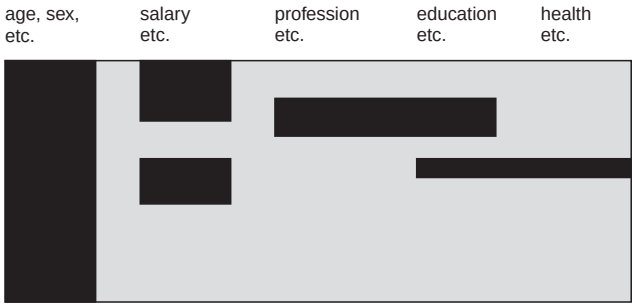
3. Mass imputation

3.1 The method

The starting point for estimating population tables based on mass imputation is the micro-database as depicted in Figure 2. Basically, the method consists of two steps. First imputing the complete set of missing data. In other words, all empty cells of Figure 2 are filled with an artificial value, the imputation. The result of this step is a completely filled micro-database: a rectangular file with a row for each person, a column for each variable, and either an observed or an imputed value in each position, see Figure 3. The observed values are still there, but the empty cells are filled with imputations, the shaded part of the figure.

In the second step population tables are estimated by adding both the observed and the imputed values. If, for example, a two-way table is requested with estimated population totals for each combination of 'age' and 'health', this table can be obtained straightforwardly by adding the observed and imputed values.
Many imputation methods have been described in the literature. The micro-database for persons consists mainly of categorical data, for which (simultaneous) hot-deck methods are usually applied. For an overview of the various imputation methods that are used in practice, see Kalton and Kasprzyk (1986).

Figure 3
A simplified version of the micro-database for persons; 'black' indicates available data, 'shaded' indicates imputed data



The main question is whether the imputation approach is a valid way to proceed from micro-databases to StatBase. In this section we shall discuss whether a consistent set of reliable estimates is obtained by this approach.

3.2 Consistency

If all estimates are obtained by adding observed and imputed values in the same fully imputed micro-database, the resulting set of estimates is numerically consistent (at least if the micro-data satisfy the edit rules). This is one of the main advantages of the (mass) imputation approach. In calculating the estimates the imputed values are treated as if they were true. A very large number of population tables can be derived (estimated) from a rectangular database and they will contain no contradictions as all estimates are based on the same (largely artificial) micro-data.

In particular, all requested small area estimates can be produced by adding observed and imputed values for all persons in the small area. If the whole population is divided into small areas and population totals are estimated for all small areas separately in this way, the sum of these estimates is equal to the estimated population total. This numerical consistency is obviously a very attractive feature.

3.3 Reliability

3.3.1 The problem

The question remains whether the estimates obtained by aggregating observed and imputed micro-data are reliable.

In section 2 reliability is made operational by requiring the population tables to be estimated either in an (approximately) design unbiased way or by a model-based method where the model used is plausible in some sense. In this section we shall argue that only a very limited number of estimates obtained by the mass imputation method correspond to an (approximately) design unbiased estimator. Reliability of the other estimates depends on the plausibility of the underlying model. It will be illustrated that in many cases this underlying model is not plausible.

By way of illustration, consider the micro-database depicted in Figure 2. The variables 'age' (100 categories), and 'municipality' (600 categories) are part of the population register and hence are known for all persons in the micro-database. The variable 'healthy or not' (two categories) is observed in the health interview survey and hence is only known for 10,000 persons. In the mass imputation approach the latter variable has to be imputed.

In order to obtain a design-based estimate, the variable 'healthy or not' should be imputed in such a way that the total number of persons that have the observed or imputed score 'healthy' is equal to the design-based estimate based on the sample; the same applies for 'not healthy'). For example, suppose that based on the sample 80% of the inhabitants of the Netherlands are estimated as feeling healthy. In that case, the imputations should be such that 80% of the observed or imputed scores are 'healthy' and 20% are 'not healthy'.

Moreover, the variable 'healthy or not' should be imputed in such a way that the relationship between this survey variable and the register variables is 'correct'. For example, for each combination of municipality and age-class, the total number of persons that have the observed or imputed score 'healthy' should be equal the respective design based estimates based on the sample. The fully imputed database should only contain information reproduced from the sample.

In the example above, there are clearly too many conditions: combining 'age' and 'municipality' leads to 60,000 classes, most of them without any observation in the survey. As a consequence, the relationship between the survey variable and the register variables in the fully imputed dataset has to be largely synthetic. For example, suppose there are no 48 year-old persons in the sample from the city of Utrecht, while there are a lot 48 year-olds in the city. For all these persons an imputation has to be constructed for the variable 'healthy or not'. After the imputations have been made, the aggregate 'total number of healthy 48 year-old persons in the city of Utrecht' can be estimated by adding the imputed values. It is clear that this aggregate is not design-based and, without further assumptions, not reliable.

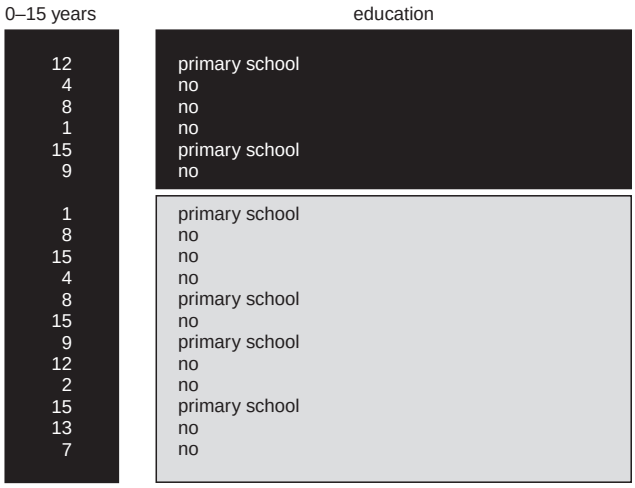
In the micro-database for persons constructed at Statistics Netherlands a number of registers are combined with a number of sample surveys. There are so many detailed register variables that it is impossible to construct imputations that lead to correct relations between all sample survey variables and all combinations of register variables. As a consequence, many aggregates that can be obtained by counting in the fully imputed database should be considered as unreliable.

3.3.2 An example

The following is an example of what can go wrong if the mass imputation approach is applied (see Figure 4). A sample survey with a variable 'completed education' is matched to a register with several variables, including 'age'. Suppose there is too much detail to render the relation between all register variables and the variable 'completed education' correct. One solution is to ensure a correct relation between 'completed education' and recoded register variables only. This recoding is done in such a way that the number of categories is reduced drastically.

The situation in Figure 4 is chosen in such a way to render the relation between 'completed education' and 'age classes' correct. One of these classes is 0-15 years. The children in this class have either completed primary school or have no completed education. In Figure 4 the imputations are carried out by hot-deck imputation within the age class. In this way, the fraction of children with 'primary school' is estimated correctly in the stated age class: six children out of 18 have the observed or imputed score 'primary school', while 12 children have the score 'no completed education'. These fractions correspond exactly to the observed fractions in the sample.

Figure 4
A survey with a variable 'completed education' is matched to a register with information about (among other things) age. In the figure only part of the data are shown. 'Black' indicates available data, 'shaded' imputations



It can be seen, however, that the relation between 'age' and 'completed education' is totally disturbed within the age class. A one-year-old child has an imputed value 'primary school' and a fifteen year-old 'no completed education'. As long as only class totals are used there is no problem in using the imputed dataset. If other aggregates are calculated based on this imputed dataset accidents may happen.

The reason that the relation between 'age' as one-year classes and 'completed education' is disturbed by the imputations, is that 'age' is not included as a factor in the imputation model. The estimate of this relation based on the imputed dataset would only be reliable if the underlying Conditional Independence Assumption (CIA) is satisfied: 'age' as one-year classes and 'completed education' are independent within the age classes as defined in the imputation model. In our example this assumption is clearly not plausible.

3.3.3 General conclusion on the reliability of estimates obtained by mass imputation

In general, when imputations have to be constructed for a sample survey variable only a limited set of auxiliary variables can be dealt with in the imputation model. The number of possible combinations of register variables is usually much higher than the sample size of the survey and hence than the number of terms that can be included in the imputation model.

After a sample survey variable has been imputed, there are two types of aggregates:

1. Totals of the sample survey variable for subsets of the population defined by variables in the imputation model. These aggregates can be estimated by adding observed and imputed values in the imputed dataset. The resulting estimates are reliable (at least if the variances are reasonably small).
2. Totals of the sample survey variable for subsets of the population that are not defined by variables in the imputation model. For these aggregates the imputation approach only generates reliable estimates if the CIA is satisfied. A loose way to formulate this assumption is: 'the elements of the subset are similar to the other elements after correction for the variables in the imputation model'.

In the example in subsection 3.3.2 the imputation model does contain the variable age class and does not contain the variable 'age' in one-year classes. As a consequence, 'the total number of children that have completed primary school in the age-class 0-15 years' is an aggregate of type 1) and can be estimated correctly by using the imputed dataset. The aggregate 'total number of eight-year old children with primary school' is an aggregate of type 2) and is not estimated reliably by this imputation approach.
In practice, the CIA is usually not satisfied. Often better model assumptions can be made specifically devised for the particular aggregate that has to be estimated. As a consequence, the mass imputation approach generates only a very limited number of reliable estimates. Note that if we were able to recalculate the imputations with a different imputation model for each new table we would lose numerical consistency.

4. Weighting

4.1 The method

The method considered in this section is based on weighting. It can be seen as a new application of traditional weighting techniques and involves four steps:

a. constructing rectangular micro-datasets from the micro-database;
b. assigning to each rectangular micro-dataset a set of weights that is derived according to some traditional weighting scheme;
c. estimating as many mutually consistent population tables of interest as possible, and
d. repeatedly reweighting the micro-datasets according to some minimal reweighting scheme for other population tables of interest.

All steps will be illustrated by means of the micro-database given in Figure 2.

4.1.1 Constructing rectangular micro-datasets

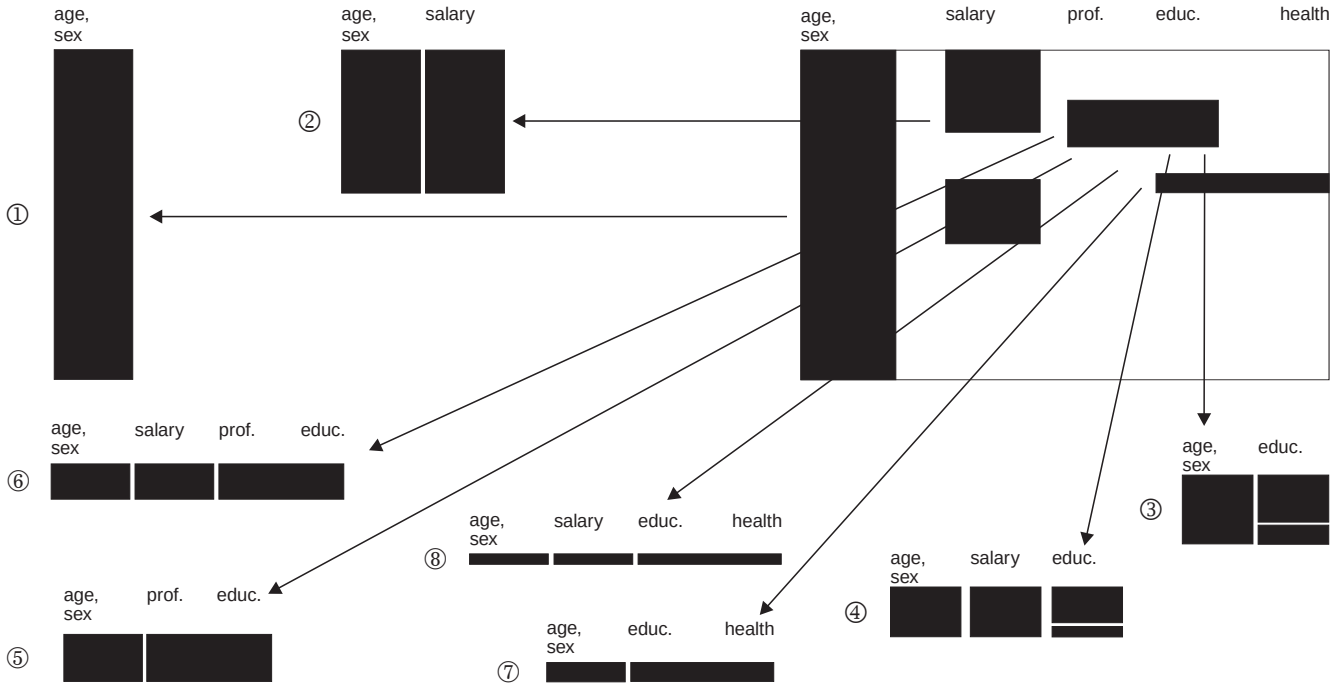
Based on the micro-database shown in Figure 2, we construct eight rectangular micro-datasets in such a way that each dataset is most suitable for estimating a specific set of population tables, see Figure 5. For example, the first dataset, is the population register, which is most suitable for estimating tables like 'age by sex by municipality'. Below, we list and briefly discuss the remaining datasets.

Dataset ② is the register with salary information enriched with information from the population register. This dataset is used to estimate aggregates like 'salary class by age'.

Dataset ③ contains a record for each person observed in either the labour force survey or in the health interview survey. It includes the variables measured in both surveys and variables from the population register are also included. This dataset is used to estimate population tables like 'completed education by age'.

Dataset ④ contains a record for each person observed in both the register with salary information and in either the labour force survey or the health interview survey. The variables included are those observed in both the labour force survey and in the health interview survey. Variables of the register with salary information and the population register are also included. This dataset is used to estimate tables like 'salary class by completed education by sex'.

Figure 5
Splitting the micro-database of figure 2 into eight rectangular micro-datasets



Dataset ⑤ is the labour force survey enriched with information from the population register. This dataset is used to estimate tables like 'occupation by age'.

Dataset ⑥ contains a record for each person observed in both the labour force survey and in the register with salary information. The dataset contains all variables measured in the labour force survey, the register with salary information or the population register. This dataset is used to estimate totals like 'salary class by occupation by age'.

Dataset ⑦ is the health interview survey enriched with information from the population register. This dataset is used to estimate aggregates like 'health by age'.

Dataset ⑧ contains a record for each person observed in both the health interview survey and in the register with salary information. The dataset contains all variables measured in the health interview survey, the register with salary information or the population register. This dataset is used to estimate totals like 'health by salary class by age'.

4.1.2 Assigning one set of weights to each micro-dataset

The next step is assigning a set of weights to each micro-dataset according to the traditional weighting approach. For each micro-dataset we derive one set of regression weights to adjust for sampling error and non-response, and to meet some (not all) consistency requirements. Utilising the general regression estimator (see Särndal et al. 1992, Chapter 6), or more generally the calibration estimator (see Deville and Särndal, 1992), for each micro-dataset we need :

1. a set of starting weights;
2. a specification of the weighting scheme (or calibration equations), i.e. a specification of the set of auxiliary variables, and
3. the population totals of the set of auxiliary variables. It should be noted that the specified set may contain auxiliary variables with population totals that are unknown, but that are estimated on the basis of earlier weighted micro-datasets. This implies that the order in which the datasets are weighted is important and should be determined carefully.

Deriving starting weights

The starting weights for the rectangular datasets given in Figure 5 can be calculated from the (net) inclusion probabilities of the original datasets (see Figure 2). If the rectangular dataset corresponds precisely to an original dataset, then the starting weights equal the inverse of the (net) inclusion probabilities. For example, the starting weights for dataset ① are identically 1, as this dataset corresponds to a complete registration. Furthermore, the starting weights for datasets ⑤ and ⑦ equal the inverse of the (net) inclusion probabilities of their corresponding samples. For dataset ③ there are several ways to derive the starting weights. One way is to derive the (net) inclusion probabilities with respect to the union of the two samples from which this dataset is derived, and subsequently to take the inverse of these probabilities. Another, more practical, way is to combine the weights of the two samples separately in some way.

To derive the starting weights for dataset ② some kind of assumption has to be made about the missing data of the corresponding register about salaries. A convenient assumption is that the missing data mechanism is independent of the missing data, conditional on some observed register variables. This assumption is called MAR, missing at random (see e.g. Gelman et al., 1995, Chapter 17). To calculate starting weights for the remaining datasets assumptions have to be made about the missing data mechanism with respect to the involved surveys in combination with the register about salaries. We shall not elaborate on this here.

Note that there are many ways to deal with incomplete registers in a micro-database. The method described in this article is to weight the observed records to compensate for the missing ones. An alternative method is to impute the missing records in the incomplete register. This method is attractive if the percentage of records missing is not too high and there is enough auxiliary information to calculate the imputations. In that case imputation of the incomplete register simplifies the estimation procedure considerably, as there are less rectangular datasets, while the danger of unreliable estimates is not too high. If the incomplete register does not contain any target variables, but is only used as auxiliary information, a third method can be applied. The population is split into two parts: the set of records for which register information is available and the set for which this information is not available. Tables will be estimated for both parts separately by constructing rectangular datasets and by weighting these. The resulting estimates can be combined afterwards. This approach is attractive if the auxiliary information is highly correlated with the target variables in the surveys.

Sequentially determining weighting schemes

After calculating the starting weights, a strategy has to be developed to weight the various datasets. An obvious strategy is to order the datasets by means of some degree of confidence in one dataset compared with another and to weight the datasets with the highest degree of confidence first. The degree of confidence may depend on indicators for several errors, such as sampling errors, non-response errors, and measurement errors. Such a strategy is important because estimated totals based on one dataset can be used as auxiliary information to weight another. For the time being and as a practical strategy, we suggest that large datasets be weighted before small datasets, as large datasets generally provide more precise estimates than small ones.

Referring to Figure 5, the most precise 'estimates' are obtained from dataset ①: aggregates like 'age by sex by municipality' can be obtained by counting in the register and have variance zero. A large number of 'estimates' of this type can be put in StatBase; all of them are reliable and the set is consistent.

Dataset ② is a good second in terms of precise estimates: the number of observed persons is much larger than that of the two surveys. The set of auxiliary variables in the weighting scheme includes as many variables from dataset ① as is allowed by the degrees of freedom. One criterion to include a variable in the weighting scheme is numerical consistency. For example, suppose 'sex' is included in the set of auxiliary variables. Then, weighted counting of the records in dataset ② leads to an estimate of the total number of men and women that is consistent with the 'estimate' based on dataset ①. As a consequence, the two-way table 'salary class by sex' is also consistently estimated with respect to sex, i.e. the marginal total numbers of men and women are equal to the number of men and women as derived from dataset ①.

The next in line is dataset ③. It can be expected that aggregates about 'completed education' can be estimated more precisely than estimates about health or occupation, as the total relevant sample size is larger. The weighting scheme for dataset ③ has to include as many estimates as possible already in StatBase. However, not all persons in dataset ③ are also observed in the salary register, and hence we have to limit ourselves to the estimates derived from dataset ①.

As the fourth candidate we choose dataset ⑤. The weighting scheme for this dataset should take into account as many relevant estimates as possible already in StatBase. These estimates are either derived from dataset ① or ③. In other words, the weights of the health interview survey are calibrated with respect to known totals in the population register and with respect to earlier estimated totals of variables like 'finished education' that are included in both surveys. This two step procedure has been proposed before in Renssen and Nieuwenbroek (1997).

The above process can be continued. In each step a different dataset is chosen and the corresponding starting weights are calibrated with respect to as many relevant estimated aggregates as possible already in StatBase.

4.1.3 Repeatedly reweighting the micro-datasets

The problem

In the previous section, it was argued that the rectangular datasets should be weighted sequentially, such that each dataset is calibrated with respect to as many estimates already in StatBase as allowed by the degrees of freedom available. The result is one set of regression or calibration weights per rectangular dataset. When using one set of weights per rectangular dataset, which can be considered as a traditional estimation method, all variables of a dataset are inflated in the same way. The main advantage of this approach is that once a set of weights has been derived, it can be applied directly to any set of variables recorded in this dataset. The resulting estimates are (approximately) design unbiased and approximate formulas exist for the design variances. Moreover, they are consistent with other estimates, provided these estimates are used in the weighting process. However, in view of StatBase, this traditional way of estimation has one obvious disadvantage.

In general it is not possible to weight a rectangular dataset taking into account all relevant estimates in StatBase. To see this, consider again the weighting of rectangular dataset ③. The dataset contains approximately 165,000+10,000 = 175,000 records. From dataset ① we can derive a table 'sex by age by municipality by marital status'. This population table has 2 * 100 * 600 * 4 = 480,000 entries. Calibrating the weights of rectangular dataset ③ with respect to this population table leads to far more restrictions than can be satisfied. There are 480,000 restrictions, while only 175,000 weights can be chosen. In general, quite similar to the problem we met when discussing the mass imputation strategy in section 3, only a limited number of auxiliary variables can be included in the weighting scheme. As a consequence, only a limited set of consistent estimates is guaranteed when assigning one set of weights to each micro-dataset. So, if we stick to the weights derived in the preceding subsection, numerical inconsistencies are inevitable. A solution to this problem is described below.

Repeated reweighting

It is argued above that in view of the consistency requirement, only a limited number of population tables can be estimated by means of one set of weights per micro-dataset. In many circumstances, this set of estimates is unnecessarily restricted. For example, suppose that the weights of dataset ③ have been calibrated with respect to 'sex by age by marital status'. This corresponds to 2 * 100 * 4 = 800 calibration restrictions, which can easily be fulfilled as there are 175,000 records. Then, the population table 'completed education by sex by age by marital status' can be estimated by using the calibrated weights and by definition this estimate is consistent with the population table 'sex by age by marital status'. However, if 'completed education by municipality' is estimated by means of these weights, the resulting estimates will be inconsistent with respect to 'municipality' obtained from dataset ① as this variable was not incorporated in the weighting process.

Still, one can force a consistent estimate by adjusting the calibration/regression weights. Namely, by recalibrating the weights with respect to the marginal distribution of 'municipality' (derived from dataset ①) and the marginal distribution of 'completed education' (estimated by means of the original set of calibrated weights). This reweighting scheme can be written as 'municipality + completed education'. Having calculated the new set of weights, the estimate of 'completed education by municipality' can be obtained by weighted counting and the result is consistent with everything else in StatBase.

This reweighting procedure is applicable for each population table to be estimated as long as the sample is large enough to estimate the table reliably in a design-based way. The reweighting scheme that is minimally required to meet the consistency requirement, is called the *minimal reweighting scheme*. Once the population tables have been estimated and these estimates added to StatBase, the adjusted weights are of no further use. Only the minimal reweighting scheme itself, i.e. the weighting model, is stored on behalf of process information.

We note that finding the minimal reweighting scheme can be quite complicated. Two or more population tables may be related because of common marginal counts such as municipality, as described in the example above. The search for related tables, and hence for minimal reweighting schemes, becomes more complicated if several marginal counts refer to the same classification variables, but at different levels. For example, the marginal count of one table refers to municipality, while the marginal count of another refers to province. Another complication arises if we take into account edit rules: 'if driving licence = yes' then 'person ≥ 18 years'. Then, when estimating population tables on driving licences, one has to take into account any estimates about age. Renssen, Kroese and Willeboordse (2000) discuss the search for minimal reweighting schemes taking into account such complications.

4.2 Consistency and reliability

The previous section described a method to obtain a set of estimates. To sum up, first rectangular datasets are constructed and for each rectangular dataset an indication is given of for which estimates will be used. A fixed set of weights is attributed to each dataset in order to obtain as much non-response correction, variance reduction and consistency as possible. Using the fixed set of weights, a large number of population tables can be estimated that are all consistent and reliable. If estimates for other population tables are required, the relevant rectangular dataset is selected and reweighted by means of the minimal reweighting scheme. Here population totals are taken into account that have been estimated and stored before. In the end, all estimates are obtained by using weights that are calibrated to known or previously estimated population totals. As a consequence, all estimates are approximately design-unbiased. Margins can be calculated and if these are reasonably small (a more or less arbitrary upper bound has to be defined), the estimates are reliable. Moreover, the estimates are also consistent. No contradictions will occur through combining estimates obtained by the weighting method as described above, provided the micro-data contains no contradictions.

There remains the question of whether the set of estimates obtained by this approach, will be rich enough to cover the need for statistical information. We claim that the richness of StatBase is determined by the comprehensiveness of the micro-database rather than by the proposed method. The reasoning behind this claim is the following. When a micro-dataset is large enough to estimate a specific table in view of the reliability requirement, then it is also large enough to be reweighted according to the minimal reweighting scheme. In other words, when a specific minimal reweighting scheme gives unstable or even undefined estimates because of insufficient (sample) cell counts, then these cell counts are also insufficient to provide reliable estimates of the table, and the related ones, directly by means of one weighting model.

5. Estimating census tables by means of the weighting approach

5.1 Introduction

The weighting approach described in the previous section has been tested at Statistics Netherlands on a prototype micro-database for

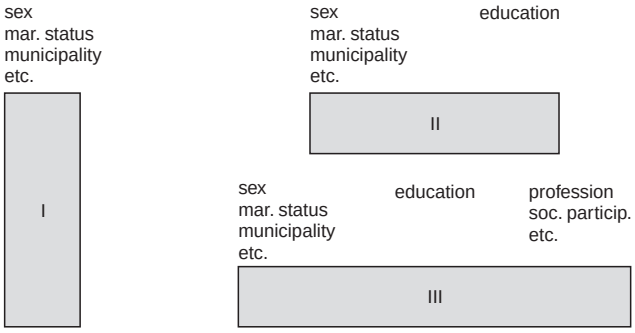
persons with data for the year 1995. A specific set of tables had to be estimated, based on the information in this micro-database. The set of tables has been derived from the Census 2001 programme.

In this section we present some preliminary results of the testing. A complete description is given elsewhere, along with the many complications encountered in practice and deliberately ignored here.

5.2 Rectangular datasets

The prototype micro-database contains information from administrative registers and surveys. Many variables in the micro-database are not relevant for estimating the census tables. The first step was to remove unnecessary variables and to recode the remaining variables in line with the specifications of the census tables. The second step was to generate rectangular datasets. As a start, three datasets were generated, see Figure 6.

Figure 6
The three rectangular datasets that are used to estimate census tables



Dataset I is the population register with variables like ‘sex’, ‘marital status’ and ‘municipality’. Other variables included are ‘nationality’, ‘age’, and ‘country of birth’.

Rectangular dataset II contains all variables that are contained in rectangular dataset I as well as the variable ‘completed education’. The records in this rectangular datasets correspond to persons observed in either the labour force survey, the health interview survey or the continuous quality-of-life survey. These sample surveys are included in the prototype micro-database of 1995 and all measure the variable ‘completed education’.

Rectangular dataset III contains all variables in rectangular dataset II as well as the variables ‘occupation’, ‘labour force participation’ and a few others. The records in this rectangular dataset correspond to the persons observed in the labour force survey.

As many census tables as possible will be estimated on the basis of these three rectangular datasets. The three datasets do not contain all information in the prototype micro-database that can be used. In particular, the available registers about jobs have not been used thus far. In a later stage of the testing this other information will be included in the estimation process.

5.3 Assigning a set of weights to each dataset

In line with the method as described in the previous section each dataset is assigned a set of weights.

The weights of dataset I are identical to one. Tables based on this dataset can be obtained by unweighted counting.

The weights of dataset II are obtained in two steps. First, starting weights are obtained by combining the original weights of the three sample surveys that contribute to the dataset. The exact combination is not described here. In the second step, the starting weights are calibrated with respect to the following weighting model:

> sex * age(5) * corop + sex * nationality + sex * age(1) + sex * marital status + sex * country of birth + age(1) * country of birth + age(5) * nationality

This weighting model was chosen to adjust for sampling errors and non-response, and to meet some consistency requirements. The variable ‘age(5)’ stands for ‘age’ recoded in five-year classes, ‘age(1)’ stands for ‘age’ in one-year classes. The variable ‘corop’ is an aggregate of the variable ‘municipality’; each *corop* consists of one or more municipalities.

The weights of dataset III are obtained by calibrating the (final) weights of the labour force survey with respect to the following weighting model:

> completed education * nationality + completed education * sex + completed education * country of birth + sex * nationality + sex * age(1) + sex * marital status + sex * country of birth + age(1) * country of birth + age(5) * nationality

Note that the weighting model contains tables with the variable ‘completed education’. In order to calibrate the starting weights, population totals are needed for these tables. These population totals are unknown, but are estimated on the basis of dataset II.

5.4 Some preliminary results

In this section some preliminary results will be presented, mainly to show that it is possible to generate a set of reliable and consistent estimates by means of the weighting approach as described in Section 4.

The tables presented in this section refer to the Dutch population aged 15-74 years in private households in 1995. The estimates are only given as an illustration, not as improved estimates of earlier published tables by Statistics Netherlands. The data and the methods used are experimental.
Estimates of the following tables will be presented:
1. ‘sex’ by ‘marital status’;
2. ‘municipality’;
3. ‘completed education’ by ‘sex’;
4. ‘completed education’ by ‘municipality’;
5. ‘labour force participation’ by ‘completed education’.

Tables 1 and 2 can be obtained immediately from rectangular dataset I and are presented below.

Table 1
An estimate of the table ‘sex’ by ‘marital status’ for the Dutch population of 15–74 years old in private households

	Sex		Total
	Men	Women	
Marital status			
Single	2,123,527	1,675,042	3,798,569
Married	3,332,038	3,392,344	6,724,382
Divorced	320,745	402,744	723,489
Widowed	79,82	344,504	424,324
Total	5,856,130	5,814,634	11,670,764

Table 4
An estimate of the table 'completed education' by æmunicipality' for the Dutch population aged 15–74 years in private households

	Completed education						Total
	Primary education	Lower secondary education	Higher secondary education	Higher vocational education	University	Not stated	
Municipality							
Aalburg	2,843	3,144	1,552	348	91	0	7,978
Aalsmeer	3,080	5,470	6,607	1,397	433	69	17,055
Aalten	2,166	5,322	5,562	506	0	0	13,555
Zweeloo	397	652	747	93	366	0	2,254
Zwijndrecht	5,155	10,220	12,200	3,613	755	0	31,943
Zwolle	10,196	17,678	29,900	13,781	4,657	0	76,212
Total	2,052,168	3,158,296	4,311,644	1,508,143	624,313	16,2	11,670,764

Table 2
An estimate of the table 'municipality' for the Dutch population aged 15–74 years in private households

Municipality	
Aalburg	7,978
Aalsmeer	17,055
Aalten	13,555
Zweeloo	2,254
Zwijndrecht	31,943
Zwolle	76,212
Total	11,670,764

Table 3
An estimate of the table 'completed education' by 'sex' for the Dutch population of 15–74 years in private households

	Sex		Total
	Men	Women	
Education			
Primary education	916,830	1,135,339	2,052,168
Lower secondary education	1,411,038	1,747,258	3,158,296
Higher secondary education	2,301,100	2,010,544	4,311,644
Higher vocational education	795,975	712,168	1,508,143
University	420,510	203,802	624,313
Unknown	10,677	5,523	16,200
Total	5,856,130	5,814,634	11,670,764

Table 3 is estimated using dataset II. Note that the set of auxiliary variables corresponding to the weighting model of this dataset contains the variable 'sex'. As a result, the set of weights can be used to estimate the table 'completed education by sex'. The marginal totals for the variable 'sex' are consistent with those in Table 1.

Table 4 also has to be estimated using dataset II. Note, however, that the set of auxiliary variables corresponding to the weighting model of this dataset does not contain the variable municipality. As a consequence, if the set of weights of dataset II were used to estimate 'completed education' by 'municipality' the table is very likely to be inconsistent with Table 2.

Table 5
An estimate of the table 'labour force participation' by 'completed education' for the Dutch population aged 15–74 years in private households

	Labour force participation			Total
	Employed	Unemployed	Economically inactive	
Education				
Primary education	508,186	99,982	1,444,001	2,052,168
Lower secondary education	1,337,661	151,554	1,669,081	3,158,296
Higher secondary education	2,680,637	195,752	1,435,255	4,311,644
Higher vocational education	1,048,997	64,710	394,436	1,508,143
University	481,814	32,786	109,713	624,313
Not stated	9,679	920	5,601	16,200
Total	6,066,973	545,704	5,058,086	11,670,764

Hence, in order to estimate this table, dataset II has to be reweighted. The reweighting model should be such that the marginal totals of the variables 'municipality' and 'completed education' are consistent with Tables 2 and 3. Note that 'corop', an aggregate of 'municipality, is part of the original weighting model of dataset 2. As a consequence, the table 'completed education' by 'corop' can be obtained consistently by using the original weights of dataset 2. In order to prevent future inconsistencies, it is wise to estimate 'completed education' by 'corop' in this way first. Subsequently, the reweighting model for dataset 2, to be used to estimate 'completed education' by 'municipality', should be such that consistency is obtained with this estimated table too.
Summarising, the reweighting model is 'municipality' + 'completed education' x 'corop'. The resulting estimate of the table 'completed education' by 'municipality' is given below. Note that this table is consistent with Tables 2 and 3.

Table 5 has to be estimated using dataset III. Note that 'completed education' is included in the weighting model for this dataset. As a result, the set of weights can be used to estimate the table 'labour force participation' by 'completed education'. The resulting table is given below. Note that the table is consistent with Tables 3 and 4.

6. Conclusions

Section 2 gives an outline of the statistical process in our institute in the near future. It is a major methodological challenge to develop an estimation technique that starts from a number of only partly filled micro-databases and results in a set of reliable estimates that are consistent and sufficiently comprehensive to satisfy user needs.

This article discusses the method of mass imputation and shows that it may lead to unreliable estimates. In section 4 a repeated reweighting strategy for estimation is described, an approach that can be seen as a new application of old weighting techniques. It is argued here that the strategy is an appropriate way to proceed from micro-databases to figures fit for publication. In section 5 some preliminary testing results are discussed.

Much work remains to be done. For example, we shall have to investigate how 'small area estimation' can be done within the framework of the method. Furthermore, more research has to be done to find out how to deal with incomplete administrative registers. And last, but not least, the method has to be tested on data from different micro-databases.

All these issues will be addressed in the near future and the resulting methods will be tested on the data described in section 5.

References

Deville, J.C. and C-E. Särndal (1992), 'Calibration estimators in survey sampling', *Journal of the American Statistical Association*, 87, pp. 376–382.

Gelman, A., Carlin, J.B., Stern, H.S., and D.B Rubin (1995), *Bayesian Data Analysis*, Texts in Statistical Science, New York, Chapman and Hall.

Kalton, G., and D. Kasprzyk. (1986), 'The treatment of missing survey data,' *Survey Methodology,* 12, pp. 1–16.

Keller, W., Bethlehem, J., Willeboordse, A., and W. Ypma (1999), 'Statistical processing in the next millennium,'*Proceedings of the XVIth Annual International Methodology Symposium on Combining Data from Different Sources, May 1999 Canada*.

Kroese, A.H., and R.H. Renssen (1999), 'Weighting and imputation at Statistics Netherlands,' *Proceedings of the IASS conference on Small Area Estimation, Riga August 1999*, pp. 109–120.

Renssen, R.H., and N.J. Nieuwenbroek. (1997), 'Aligning estimates for common variables in two or more sample surveys,' *Journal of the American Statistical Association ,* 92, pp. 396–374.

Renssen, R.H., Kroese, A.H., and A.J. Willeboordse (2000), 'Aligning Estimates by Repeated Weighting', forthcoming.

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

The use of administrative registers to reduce non-response bias in household surveys

Linda Geuzinge, Johan van Rooijen and Bart F.M. Bakker

1. Introduction

Non-response is a serious problem in social statistics and the underlying sample surveys. It makes sample survey estimates questionable, because of a potential bias that is hard to measure. Non-response rates in sample surveys of Statistics Netherlands have increased sharply in the last decades, prompting a growing concern about the problem of bias in recent years. In 1977 the Labour Force Survey (LFS) had a response of approximately 90%, while for the same survey in 1995 response was only 60%. For the quality-of-life surveys, the response rates declined from 72% in 1974 to 50% in 1995. Most non-response consists of refusals: approximately 65%. An international study on sample survey non-response showed that Dutch response rates are very low compared with other countries. For instance, the Dutch LFS achieved a response rate of 60% in 1995, while the response rate for LFS's in other European countries varied between 82% and 93% that year (De Heer, 1996).

Statistics Netherlands has taken several initiatives to reduce non-response, many of which are concerned with how sample surveys are presented to the public (e.g. introduction letters), interviewer monitoring and the use of an optimum combination of different data collection methods (e.g. Vousten & De Heer, 1998). These efforts have reduced the overall non-response rate to some extent. The question remains whether a high response rate implies unbiased results. As the non-response error is a function of the non-response rate and the difference between average scores of non-respondents and respondents, increasing response rates do not always reduce the non-response error (Groves, 1989).

Improving response rates might not be enough; indeed, this might even increase the non-response bias. This is the case if the refusals differ substantially from the non-refusals in the average scores on the target variables of the sample survey, and improvement of the fieldwork leads to more response from the non-refusals in the non-contact group. As a result, the average scores of the respondents differ more from the non-respondents than before, because the non-contact group is composed more homogeneously of refusals.

Reducing the non-response bias must be the ultimate goal. This bias does not occur when non-response is random. The main non-response category consists of refusals. Refusing to co-operate with a sample survey implies a more or less conscious decision and therefore leads to a selection process based on the interest in the survey topics. This will lead to serious bias, in particular with respect to the target variables of the sample survey.

Weighting by post-stratification can reduce the non-response bias. Up to now, Statistics Netherlands has only made use of the information from the population registers to weight sample survey data. These registers contain population information on age, sex, marital status and urbanisation. The estimates from the weighted sample surveys will be improved if these characteristics are associated with the chance of response and correlate with the target variables of the survey, assuming that the respondents in the cells of the cross-tabulation of these variables are representative.
In this article we shall present an elaborated version of the method of weighting by post-stratification, consisting of the following steps. Firstly, administrative registers with data on the entire population are exactly matched with each other and with the response and non-response of the particular survey. We expect the variables in the registers to have a greater correlation with the target variables in the sample survey than the variables form the population register. Secondly, we present a method to select variables from the linked registers that correlate strongly with the target variables in the sample survey and are associated with the chance of response. Thirdly, we apply a previously developed method for weighting data to univariate and bivariate instead of multivariate distributions. Since the small size of the sample in question seriously limits the maximum complexity of the employed weighting model, this method enables unnecessarily complex components of the weighting model to be avoided, and thus additional, different components to be included in the weighting model.

We shall present an empirical example in which we weight the response of the 1995 Netherlands' health interview survey (HIS) with the use of administrative registers on jobs and social security benefits. In particular we shall examine the effects of the new weighting procedure on the estimation of the number of bed-days spent in hospitals. In accordance with the idea that refusing to respond to a sample survey leads to a selection process based on the interest in the topics of the survey and thus to serious bias, we expect that more people with health problems and a high medical consumption will respond to a health interview survey than others. This leads to the hypothesis that medical consumption will be overestimated in the HIS.

In addition to the results for bed-days, we shall estimate the effects of the new weighting procedure for educational level. It is often assumed that higher educated people are more willing to respond to sample surveys than those with lower education, because they are more convinced of the usefulness of surveys and are more loyal to the government. However, as it is not possible to weight on educational level, as there is no integral register on education, we could only collect indirect evidence for a selectivity bias by educational level: we estimate the selectivity on variables that are registered integrally and correlate strongly with educational level. If the new weighting procedure corrects for this selectivity, it should lead to a lowering of educational levels.
Section 2 presents the data: the HIS and the administrative registers used. Section 3 describes the developed method and its application to the HIS. The last section discusses the results and evaluates the usefulness of the developed method for reducing non-response bias.

2. The Netherlands health interview survey and administrative registers

2.1 The health interview survey

The household sample survey used in this study – the Netherlands health interview survey (HIS) – aims to give a view of the health status and medical consumption of the Dutch population in relation to its background characteristics. We used the 1995 HIS for the non-response analysis.
The information is collected by means of interviewing a representative sample of the population in the Netherlands. For practical reasons the sample survey is restricted to the non-institutionalised population. As the survey aims to provide information on a household as well as an individual level, an address sample is used. No more than three households at the same address and no more than four persons per household are

interviewed. If a household consists of more than four members, the core of the household (i.e. the 'head' of the household and his/her partner) is interviewed, followed by two (or three if there is no partner) randomly selected additional members (usually the children). Where necessary proxy interviews are used. The HIS is a continuous survey, so interviews take place during the whole year. In 1995, 6,643 addresses were sampled, with a total 17,522 residents. Some of these people were justified non-respondents as they were excluded by the sample design, for example because they were the fifth person in the household. Of the 17,522 residents, 9,930 were actually interviewed. The non-response rate amounted to 41% in 1995, and 61% of this non-response consisted of refusals.

2.2 The administrative registers

The following administrative registers were used:
- The population register (PR);
- The administration of employee insurance schemes, jobs (AEIS-jobs);
- The administration of employee insurance schemes, unemployment benefits (AEIS-UB);
- The administration of employee insurance schemes, disablement benefits (AEIS-DB);
- The annual survey on employment and earnings (ASEE);
- The administration of public sector employees disablement benefits (APEDB);
- The social assistance benefits administration (SABA).

All these sources refer to 1995. The main characteristics of the registers are described below.
The population register covers everyone in the Netherlands officially registered in his or her municipality of residence. The PR contains data on demographic characteristics like sex, age, marital status, country of birth and number of children.

The AEIS-jobs contains integral information on employees who had a job in the private sector in the year in question with the exception of self-employed people. The target population of the ASEE covers all jobs in the Netherlands except those of the self-employed. A second difference with the AEIS-jobs is the fact that the ASEE is partly a sample survey, whereas jobs in the public sector are covered integrally. In addition, the ASEE only contains data about jobs on 31 December of the year concerned. People included in the AEIS-jobs or in the ASEE had one or more jobs in the course of 1995. These files also contain information on characteristics of these jobs, for example when they started and ended, whether the employee was a regular, temporary or a stand-by employee, some information on wages, the most important economic activity of the company and the company size.

The AEIS-UB, AEIS-DB and SABA contain information about social benefits. The AEIS-UB registers people from the private sector who are entitled to some form of social benefit in the year concerned. The AEIS-DB represents all people in the private sector entitled to a benefit for disablement, except the self-employed. The benefits for the disabled from the public sector are available from the APEDB. The SABA represents all people who are entitled to social welfare at the end of the year concerned. People who appear in one of the social benefit administrations have received one or more benefits in the course of 1995. These registers also contain information on the type and period of the benefit, the official body paying it and some background characteristics of the claimants.

3. The developed method for reducing non-response bias

3.1 An overview

Non-response occurs when a population element that has been selected in the sample is not interviewed and thus the desired information is not obtained. Such elements are indicated with the term 'non-respondents'. This leads to a separation of the data into two parts, the response part and the non-response part. The gross sample is the initial sample and exists of all sample elements; it contains the response and the non-response. The net sample consists of all respondents and also contains the survey information. The gross sample and the net sample will be matched with administrative registers, creating the linked gross and net samples.

The method developed to reduce non-response bias in household sample surveys consists of the following steps:
- selection of the most important target variables from the household sample surveys;
- matching micro-level data from the net sample, the gross sample and some integral registrations;
- selection of the potential weight variables. From the administrative registers, we select the variables which we expect to have a strong correlation with the target variables of the household sample surveys, and on which we expect selectivity for the response probability. These selected variables are called the potential weight variables. The association between the potential weight variables and the target variables and between the potential weight variables and the response probability is verified empirically, resulting in an order of the importance of the potential weight variables;
- for smaller samples it is not possible to include all potential weight variables in one multi-dimensional weighting scheme, because of the occurrence of empty cells and the risk of an unstable weighting solution. Therefore, the dimensionality of the employed cross tables was reduced as much as possible, while simultaneously taking into account the multivariate associations between the potential weight variables;
- The results of the new weighting procedure will be calculated and compared with the results of the original weighting procedure. One of the problems that occurs during the matching of the gross and the net sample of the household surveys with the administrative registers is that a number of records in the response of the household sample survey are not matched. In the last phase these records are added to the file and receive the weight of the original weight variable and as such are included in the calculation of the final result.

3.2 Selection of target variables from the household survey

The correlation with the target variables from the sample survey is an important criterion for the selection of the potential weight variables. Sample surveys usually contain many target variables, and this is also true of the HIS. In this study it was not necessary to examine the correlations between all these target variables and the potential weight variables. From the point of view of manageability, it was necessary to choose from the target variables, and for this choice the main criteria were the social significance of the data, the importance of the variables for publications and whether the selected variables cover the fields of study. If a study covers several fields, the most important indicators will be chosen.

For a manageable number of target variables a selection from the HIS was necessary. In this case the choice was based on the importance of the variables in publications on the health status and medical consumption (Van Baal, 1997; SN, 1996a; SN, 1997). The following indicators were chosen:

Indicators for the health status:
- the subjective health status of the interviewed person/perceived health status ;
- number of chronic conditions;
- number of physical complaints ;
- number of physical disabilities ;
- Quetelet-index: a measure for overweight and underweight.

Indicators for medical consumption:
– general practitioner (GP) consultations;
– Specialist consultations;
– hospital admissions;
– use of prescribed medicines;
– use of non-prescribed medicines;
– physiotherapist consultations.

Factor analysis was used to test whether these selected variables are indicators for the health status and medical consumption. Factor analysis is an important tool for the selection of the most important indicators. It searches for latent, not measured variables that explain the collective variance of the variables which are actually measured in the survey. The measured variable with the highest factor loading is expected to be the best single indicator for the collective variance of all measured variables which contribute to the same factor. Once variables have been selected which belong to a certain field of research, factor analysis can establish which variables are the best indicators for a certain field.

We applied the so-called split-half method. For the first half of the net sample factor analysis was used to search for the best fitting model for the eleven selected target variables. Next, this model was tested on the second half of the net sample to ensure that the established factor structure was not a 'chance hit'. The fit of the model is given in the following table.

Table 1
The fit of the model for health status and medical consumption

	χ^2	Df	Agfi [1)]	P	N
Model first half	404	34	.97	.00	4 936
Model second half	559	34	.96	.00	4 994

[1)] Agfi: the Adjusted goodness of fit index

Taking into account the number of respondents, the model fits the data well for the first half of the net sample. After testing the model on the second half, it can be concluded that the model has a reasonable fit. The results were obtained by using LISREL VI (Jöreskog & Sörböm, 1986). The best fitting model contains three factors; the first and the second describe health status and medical consumption respectively. The third factor has only a high factor loading for the variable 'use of prescribed medicines'. The conclusion is that the selected variables are indeed indicators for the health status and medical consumption. These variables will therefore be called the target variables.

3.3 Composing the gross sample and linking the administrative registers, gross and net sample

The next step was to compose the linked gross sample, i.e. to select the non-respondents and respondents from the population register. The initial gross sample consists of all the sampled addresses, therefore the measurement unit is the address. The linked gross sample represents all the people living at the addresses in the initial gross sample. The measurement unit of the linked gross sample is the individual. The linked gross sample is assumed to be a correct representation of the target population of the HIS.

Having composed the linked gross sample, the data were exactly linked with the net sample and the integral registrations at the level of the individual. In this way a database was created with information on the response and the non-response from several different integral registrations. More in particular, the database contained more information on non-response than the weighting frame: the population register. In addition to demographic characteristics, information about jobs and social security payments was also used.

When the records from the net and gross sample had been linked with corresponding records from the integral registers, variables from the registers had to be selected as potential weighting variables. A selection had to be made because the size of the sample restricts the complexity of the applied weighting scheme. In other words, if too many variables from the registers are added to the weighting scheme, this scheme, and therefore the weighting solution, would be complicated to such an extent that a stable weighting solution would not be attained. The number of variables that can be added to the weighting scheme decreases with decreasing sample size. As the sample size of the survey concerned was intermediate, the above-mentioned constraints were expected to limit the number of variables that could be included in the weighting scheme.

As variables had to be selected from the registers, a sound criterion was needed for the selection. Applying this criterion would have to result in selection of variables that would correct optimally for the bias caused by selective non-response. In fact, a variable from the registers has to satisfy two criterions in order to be relevant for the weighting procedure. Firstly, it has to correlate with the probability of response. If this correlation exists, the sample will deviate from the population with regard to this variable. Secondly, it has to correlate with the target variables in the sample, thus ensuring that a weighting procedure based on the variable in question will also decrease the bias of the sample with regard to the target variables.

The following potential weight variables were selected: sex, age, marital status, family type, position in the family, number of children, ethnic group, source of income, gross annual wage, income at the address and type of employee. An overview of these variables and their additional categories can be found in appendix 1.

For the selection of potential weighting variables, we searched the literature for relevant ones. The same sorts of reasons about the expected correlation between the potential weight variables and the target variables can be argued for all the selected potential weight variables. *Age* was selected because older people more often have bad health and higher medical consumption than younger people (SN, 1997).

According to Statistics Netherlands (1998), *ethnic minorities* suffer more from illness than Dutch-born people. However, the barriers for the consumption of medical services are relatively high, so we expect slightly lower medical consumption.

We expected a high correlation with health status and medical consumption for the potential weight variables *source of income* and *household income level*. In general socio-economic status (in terms of educational level, occupational prestige, wages or income) correlates negatively with health status and positively with medical consumption (e.g. Cavelaars et al., 1998a & 1998b, Van de Mheen, 1998). Furthermore, we expected people in employment to be better off than those on unemployment or disablement benefit. There are two reasons for this: first, unemployment and disablement lead to a sharp fall in income; and second, unemployment and disablement are associated with social isolation and a low status in society. Both phenomena will lead to a higher health risk and therefore higher medical consumption. (Batenburg, Smeenk & Ultee, 1995; Van de Mheen, 1998).

The potential weight variables also have to be selective for the response. For the variable 'age', for example, we expect older people to be home more often than younger people, so the chance of response will be higher. On the other hand, ethnic groups have lower response rates because of language problems. The response rate is also relatively low in neighbourhoods inhabited by predominantly lower income groups.

3.4 The selection of potential weighting variables

The above-mentioned correlations between the variables from the registers and the chance of response were determined by means of logistic regression analysis. This is the most suitable technique as responding or not responding is represented by a dichotomous variable. The fact that many of the variables from the registers were nominal variables presented no problem as these were automatically transformed into dummy variables by the logistic regression procedure. The correlations between each of the variables from the registers and the target variables from the sample survey were determined by means of multiple regression analysis. In these analyses the variables from the registers were the dependent variables whereas the target variables from the sample survey were the independents. As only the correlation was relevant, the exchange of dependent and independent variables was not a problem. In the case of nominal variables, an optimal scaling procedure based on canonical correlation analysis was used to establish new category quantifications. The application of these new category quantifications resulted in the transformation of these variables into quasi-ordinal variables that could subsequently be used in the multiple regression analyses.

Both the multiple regression analyses and the logistic regression analyses resulted in correlation coefficients. Thus, for each variable form the registers two correlation coefficients were determined. These represented the correlation with the chance of response and the correlation with the target variables from the sample survey respectively. As stated above, both correlations must deviate substantially from zero for a variable to be important for a weighting procedure. Therefore, the pairs of correlation coefficients for each variable from the registers were multiplied and the product of correlation coefficients was used as a measure of the significance of each variable for a weighting procedure. In this way, the two criteria which must be met by a variable from the registers were united in one objective measure.

The correlations with the chance of response and with the target variables from the sample survey respectively and the product of these correlation coefficients are presented in Table 2. The reason for using the product of the two correlation coefficients is that if the correlation of a variable from the registers with the chance of response is strong, the sample will deviate substantially from the population on this variable. Subsequently, the inclusion of this variable in a weighting scheme will in principle lead to a large correction of the bias present in the sample. However, if the correlation of the variable concerned with the sample survey target variables is zero, the correction of the bias due to weighting will not include the target variables. In this case, the product of correlation coefficients is zero, indicating unequivocally the lack of 'corrective power' of the variable in question. In the reverse case of strong correlation with the target variables and zero correlation with the chance of response, an analogous argumentation applies. The only drawback of this construct is the fact that the distribution of this statistic is unknown and therefore, levels of significance cannot be established. However, as is often the case with relatively large samples, practically all examined correlations will turn out to be highly significant. Therefore, the importance of significance levels is reduced when using relatively large samples, whereas the importance of the magnitude of the correlations is increased.

The next step is to determine the weighting scheme and carry out the weighting procedure. The objective of the weighting scheme is to divide the population into categories. The weighting procedure subsequently assigns weights to each category in the sample. As a result, the weighted sample will be representative of the population with regard to the categories used, thus eliminating the initial bias caused by selective non-response. The application of a weighting scheme corresponds with post-stratification. Evidently, the ideal stratification would simply divide the population into as many categories as possible, enabling the most complete weighting of the sample. This maximum stratification would be obtained by including all the available weighting variables in a multi-dimensional cross table, of which each cell would then represent a stratum. The cross table based on the linked gross sample is used as the frame of reference as the linked gross sample is assumed to be representative for the population. As a result of selective non-response, the cross table based on the net sample will deviate from this frame of reference. The objective of the weighting procedure is to assign such weights to the records of the net sample that the cross table based on the weighted net sample will exactly reproduce the cross table based on the linked gross sample.

In practice, the above-mentioned multi-dimensional cross table cannot be used as a weighting scheme. Because the HIS sample is relatively small, the cross table contains many empty cells in the net sample. Therefore, in many cases a cell in the net sample will be empty whereas the corresponding cell in the gross sample will be filled. In these cases it is not possible to assign weights to records of the net sample to fill these empty cells, as at least one case is

Table 2
Correlations of each potential weighting variable with the target variables from the Netherlands health interview survey (HIS) and with the chance of response [1)]

Potential weighting variables	from register	Correlation with		Product of correlations
		target variables	response chance	
		R1	R2	R1*R2
Age	PR	.57**	.09**	.051
Position in family	PR	.34**	.12**	.041
Marital status	PR	.38**	.07**	.027
Family type	PR	.15**	.13**	.020
Ethnic group	PR	.13**	.10**	.013
Number of children	PR	.18**	.06**	.011
Source of income	AEIS-jobs/AEIS-UB/AEIS-DB/ ASEE/APEDB/ SABA	.17**	.06**	.010
Income at address	Ditto	.27**	.03**	.008
Type of employee	AEIS-jobs/ASEE	.21**	.03**	.006
Gross annual wages	AEIS-jobs/ASEE	.28**	.02**	.006
Sex	PR	.15**	.01	.002

Level of significance:
* = p <.05
** = p >.01

1) The variables are ordered according to the product of the correlations.

needed to assign a weight to. As a consequence, the cross table from the gross sample cannot be reproduced by the net sample by means of weighting. In addition, the net sample will contain many nearly empty cells, causing the weighting solution to become unstable as the weighting program used determines such weights that the cross table of the gross sample is reproduced exactly by the weighted net sample. This is so demanding that very large or even negative weights will arise, which is an undesirable solution. As a consequence of these constraints, the complexity of the weighting scheme, i.e. the number of cells, needs to be reduced. However, at the same time the 'corrective power' of the weighting scheme has to be kept at a high level and therefore the weighting scheme has to be optimised. The strategy used to optimise the weighting scheme is described below.

The first criterion for optimisation of the weighting scheme consists of the product of correlations (see Table 2) as these are a measure of the correction of the bias that will be attained when the variable in question is included in a weighting scheme. Therefore, the variables from the registers were prioritised for inclusion in the weighting scheme according to the product of correlations. However, the variables could not simply be added successively to the weighting scheme according to their priority, since the mutual correlations had not yet been taken into account. That is, if there are correlations between the variables themselves, these variables have a high degree of overlap in their information content. As a consequence, once one of these variables has been selected, the remaining variables will not represent much added value. Therefore, a factor analysis was performed on the variables from the registers in order to establish the pattern of mutual correlations. For each of the relevant factors a variable was selected that would subsequently represent the factor in question. The criterion for this selection consisted of a combination of a high factor loading and a high product of correlations. Thus, the selection criterion ensured the representativeness of the variable for the factor and the power of the variable to correct for the bias. Five factors were found and the variables 'gross annual wage', 'age', 'type of family', 'ethnic group' and 'sex' were selected.

The selected variables were included in multi-dimensional cross tables, as this would generate the most complete post-stratification and consequently the most complete weighting. However, 3-dimensional cross-tables were used at the most, since empty or nearly empty cells arose when cross tables with more dimensions were used. In addition, the variable 'sex' was not included in the 3-dimensional cross tables as the weighting scheme became too complex and this variable was the least relevant for correction of the non-response bias (see Table 2). The variable 'sex' was added to the weighting scheme separately. Next, loglinear analysis was carried out to determine whether higher order interactions were present. The objective of this analysis was to determine whether the 3-dimensional cross tables could be reduced to 2-dimensional cross tables or frequency tables without loss of information. When higher order interactions are absent, the 3-dimensional cross table does not contain more information than the underlying 2-dimensional cross tables or frequency tables. However, the 3-dimensional cross-table does contain more cells, which makes the weighting model as complete but less complex when 2-dimensional cross-tables or frequency tables are used. Only in the cross table 'age' x 'ethnic group' x 'type of family' was the third order interaction found to be relevant. However, this cross table contained empty cells that were not empty in the gross sample, which necessitated the use of the underlying 2-dimensional cross tables. The remaining 3-dimensional cross tables could be reduced to their underlying 2-dimensional cross-tables as no third order interactions were relevant. All second order interactions played a significant role. Therefore, these cross tables were included in the weighting scheme except for 'age' x 'ethnic group' and 'age' x 'gross annual wage' as these also contained empty cells.

The described strategy results in an optimum weighting procedure which uses as few strata as possible while simultaneously maximising the power to correct for the bias which had been caused by selective non-response. The assignation of weights to the records from the net sample was done by means of the program Bascula 3.0, developed at Statistics Netherlands.

The remaining variables from the registers were not included in cross tables, but were introduced separately in the weighting scheme. As a consequence, the frequency distributions of these variables in the linked gross sample are reproduced by the weighted net sample. The importance of this step lies in the fact that Statistics Netherlands regularly tabulates by these variables. Since Statistics Netherlands strives to publish only one figure for one phenomenon (the so-called one-figure concept), figures need to be tabulated by variables that are calibrated in such a way that the marginal distributions of these variables are reproduced. The final weighting scheme is described in Table 3.

Table 3
The weighting scheme for the Netherlands health interview survey (HIS)

Age x family type
Ethnic group x gross annual wage
Gross annual wage x family type
Family type x ethnic group
Type of employee
Position in the family
Number of children
Income at address
Sex
Marital status
Source of income

3.4 Results after weighting

The concrete result of the weighting procedure consists of a variable with weights that has been added to the net sample of the HIS. In order to obtain an indication of the effect of the new weighting procedure, the variables 'number of bed-days in hospital' and 'level of education' were used. The net sample was weighted with the new and former weighting variable respectively, and the population totals for the selected variables were calculated.

The new weighting procedure for the HIS resulted in slightly different population totals compared to the former weighting. The estimate of the total number of bed-days in hospital came out slightly lower than the original estimate: 418 thousand. The difference is small with regard to the level of education. With the new weighting variable, a trend towards a smaller number of higher educated people can be detected, although this difference is small. The results are shown in Table 4.

Table 4
Results of the new weighting procedure for bed-days in hospital and level of education

	Weighted with	
	Original weighting variable	New weighting variable
	x 1,000	
Number of bed days in hospital	11,610	11,192
	%	
Level of education		
1. primary education	17.4	17.6
2. lower secondary education	28.4	28.7
3. higher secondary education	34.6	34.7
4. tertiary education	19.6	19.0

4. Discussion

Non-response is a serious problem in social statistics and social sample surveys. It makes sample survey estimates questionable because of a potential bias that is difficult to measure. Non-response rates in sample surveys of Statistics Netherlands have increased sharply in the last decades, leading to a growing concern about the bias effect. Response rates for the Netherlands are very low compared with other countries. Bias does not occur when non-response is random. Most non-responders are people who refuse to comply with the sample survey, which implies a more or less conscious choice and therefore leads to a selection process based on the interest in the topics of the sample survey. This will lead to serious bias, in particular on the target variables of the sample survey.

Weighting by post-stratification can reduce non-response bias. Up to now, Statistics Netherlands has only made use of the information from the population registers to weight sample survey data. These registers contain population information on age, sex, marital status and urbanisation. The estimates from the weighted sample surveys will be improved if these characteristics are associated with the probability of response and correlate with the target variables of the sample survey, assuming that the respondents in the cells of the cross tabulation of these variables are representative.

This paper describes the test of an elaborated version of the method of weighting by post-stratification, containing the following steps. Firstly administrative registers with data on the whole population are exactly matched with each other and with the response and non-response of the sample survey concerned. We expect the variables in the registers to correlate more strongly with the target variables in the sample survey than the variables from the population register. Secondly, we developed a method to select variables from the linked registers that correlate strongly with the target variables in the sample survey and are associated with the probability of response. Thirdly, we applied an already developed method for weighting data to univariate and bivariate instead of multivariate distributions. Since the small size of the sample in question seriously limits the maximum complexity of the weighting model used, this method enables unnecessarily complex components of the weighting model to be avoided, and additional, different components to be included in the weighting model.
We tested the method by weighting the response of the 1995 Netherlands health interview survey (HIS) with the use of administrative registers on jobs and social security benefits. In particular we examined the effects of the new weighting procedure for the estimation of the number of bed-days spent in hospitals and level of education.

In accordance with the idea that refusing to respond to a sample survey leads to a selection process affected by the interest in the topics of the sample survey and will therefore lead to serious bias, we expected more people with health problems and high medical consumption to respond to a health interview survey than others. This led to the hypothesis that the medical consumption will be overestimated in the HIS. The hypothesis is not rejected by our empirical results. Applying the new weighting scheme leads to a reduction in the estimation of the number of bed-days in hospitals of 418 thousand, approximately 4%. Of course it is still uncertain whether the non-response bias has been reduced by this procedure, as we do not know what the 'right' figure is. Nevertheless, the confidence in the applied method has grown, because the hypothesis was not rejected. Further research is necessary to generalise the conclusions.

One increasingly expressed statement on non-response bias is that the response is selective according to socio-economic variables, in particular educational level. The results of our research show that there is some response bias on socio-economic variables like source and level of income, wages and type of employee, but that the probability to respond differs only slightly according to these variables. Demographic variables like, age, position in the family, marital status, family type, ethnic group and number of children are much more important.

Unfortunately, we were not able to test the selectivity on educational level, as we do not have integral data on this variable. However, the new weighting procedure did not lead to significantly different estimates for educational level. Furthermore, the educational level correlates fairly strongly with the other socio-economic variables. Therefore, we assume that the selectivity on educational level will not be high.

The employed method consists of several steps that are carried out sequentially. Although the individual steps constitute sound methods, one might wonder whether the overall algorithm is optimal. Therefore, further research is required to determine whether the algorithm as a whole can be improved. For this purpose, simulation studies using this algorithm would be suitable; it would for instance be interesting to integrate the method of reducing non-response bias with the method of estimation developed by Kroese et al. (2000). The method of reducing non-response bias gives a good indication of the variables that have to be used in the early stages of the repeated weighting models.

The value of the developed method for reducing the non-response bias in sample surveys depends on the availability and the quality of the integral administrative registers. In particular we expect that the method will significantly reduce the non-response bias in health interview surveys if the available administrative registers relate to health or medical consumption. For instance a register on the use of GP services would be very useful. Nevertheless, we hope that we have convinced the reader that the developed method is a promising road to reduce the bias in sample surveys caused by non-response.

References

Baal, M. van, 1997, Regionale gezondheidsverschillen in Nederland, 1989/1995 (Regional health differences in the Netherlands, 1989/1995), In: *Maandbericht Gezondheid* vol. 1997, no. 3, pp. 5–7.

Batenburg, R.S., W.H. Smeenk en W.C. Ultee, 1995, De sociale verdeling van werkloosheid: een werkloze onderklasse? (The social stratification of unemployment: an unemployed underclass?) In: J. Dronkers en W.C. Ultee (eds.), *Verschuivende ongelijkheid in Nederland: Sociale gelaagdheid en mobiliteit* (Assen: van Gorcum), pp. 266–283.

Cavelaars, A.E.J.M., A. E. Kunst, J.J.M. Geurts, R. Crialesi, et al, 1998a, Differences in self-reported morbidity by educational level. A comparison of 11 western European countries, In: *Journal of Epidemiology and Community Health*, vol. 1998, no. 52, blz. 219 e.v.

Cavelaars, A.E.J.M., A. E. Kunst, J.J.M. Geurts, U. Helmert, et al, 1998b, Morbidity differences by occupational class among men in seven European countries. An application of the Erikson-Goldthorpe social class scheme, In: *International Journal of Epidemiology*, vol. 1998, no. 27, pp. 222 e.v.

De Heer, W., 1996, *International response trends: description and explanation*. (Rome: Paper presented at the 7th Workshop on Household Survey Non-respons, 2-4 October 1996, ISTAT).

Groves, R.M., 1989, *Survey errors and survey costs* (New York: Wiley).

Jöreskog, K., & D. Sörböm, 1986, *LISREL VI. Analysis of linear structural relationships by maximum likelihood. Instrumental*

variables and least square methods (Mooresville Indiana: Scientific Software).

Kroese, A.H., R. Renssen & M. Trijssenaar, 2000, Weighting Or Imputation: Constructing A Consistent Set of Estimates Based on Data from Different Sources, In: P.G. Al & B.F.M. Bakker (eds.), Re-engeneering social statistics by micro-integration of different sources. This issue of *Netherlands Official Statistics.*

Statistics Netherlands, 1996, *Netherlands Health Survey 1981–1995* (The Hague: SDU publishers/Statistics Netherlands).

Statistics Netherlands, 1997, *Vademecum Gezondheidsstatistiek 1997 (Vademecum of health statistics 1997)* (The Hague: publishers/Statistics Netherlands).

Statistics Netherlands, 1998, *Allochtonen in Nederland 1998* (*Foreigners in the Netherlands 1998)* (Voorburg: Statistics Netherlands).

Van de Mheen, D., 1998, *Inequalitites in health, to be continued? A life-course perspective on socio-economic inequalities in health* (Rotterdam: PhD dissertation Erasmus University Rotterdam).

Vousten, R., & W. De Heer, 1998, Reducing non-response: the POLS fieldwork design, In: B.F.M. Bakker & J.W. Winkels (eds.), Integration of household surveys. Design, advantages and methods. In: *Netherlands Official Statistics,* vol. 13, summer 1998, pp. 16–19.

Appendix
The used target and potential weight variables for the HIS analyses

Target variables from the HIS

The subjective health status of the interviewed person/perceived health status
1. very good
2. good
3. moderate
4. sometimes good and sometimes bad
5. bad

Number of chronic conditions
The number of times somebody has a chronic condition (list of 24 chronic conditions)

Number of physical complaints
The number of times somebody answered 'yes' on a list of 23 stress-related complaints (VOEG)

Number of physical disabilities
The number of times somebody has extreme difficulty with a certain activity or finds it is impossible to do (list of 7 activities of the OECD)

The Quetelet-index: a measure for overweight and underweight
1. Underweight
2. Normal
3. Overweight

General practitioner (GP) consultations
1. Yes
2. No

Specialist consultations
1. Yes
2. No

Hospital admissions
1. Yes
2. No

Use of prescribed medicines
1. Yes
2. No

Use of non-prescribed medicines
1. Yes
2. No

Physiotherapist consultations
1. Yes
2. No

Sex
1. Man
2. Woman

Age
1. 0–11 years
2. 12–17 years
3. 18–24 years
4. 25–34 years
5. 35–44 years
6. 45–54 years
7. 55–64 years
8. 65–74 years
9. 75 years or older

Marital status
1. Unmarried
2. Married
3. Widowed
4. Divorced

Family type
1. Married couple with children
2. Man with children
3. Woman with children
4. Cohabiting couple with children
5. Persons not living in a family

Position in the family
1. Man in a married couple without children
2. Man in a married couple with children
3. A single person with one or more children
4. Woman in a married couple
5. Child
6. Single person without children
7. Man or woman who live together with one or more children

Number of children
1. No children
2. 1 child
3. 2 children
4. 3 children
5. 4 children
6. 5 or more children

Ethnic group
1. Native Dutch
2. Non-Dutch belonging to one of the policy 'target groups'
3. Non-Dutch not belonging to one of the policy 'target groups'

Source of income
1. Labour
2. Disablement benefit
3. Unemployment benefit
4. Social security benefit
5. No labour or benefit
6. Labour and one or more benefits
7. Several benefits

Gross annual wage
1. None or unknown
2. 1– 30,775 Dutch guilders
3. 30,776–400,000 Dutch guilders

Income at address
1. None or unknown
2. 1– 25,000 Dutch guilders
3. 25,001– 50,000 Dutch guilders
4. 50,001– 75,000 Dutch guilders
5. 75,001–400,000 Dutch guilders

Type of employee
1. Employee with permanent position
2. Flexible worker
3. Working on an employment scheme
4. Not employed

One figure for the supply and demand of services

Bart F.M. Bakker and Johan van Rooijen

1. Introduction

One of the advantages of the SSB approach is that a large part of the integration process that up to now has been done at macro-level, will be performed at micro-level. However, macro-integration will still be necessary, for example for consistency with the national accounts or with the outcomes of business surveys. This article describes the macro-integration process of business and household surveys on the quantities of services within the frame of the so-called supply-and-demand matrices. In this integration process several adjustment factors are estimated. After the determination of the integrated quantities of services, it is necessary to achieve consistency between the macro-totals and the aggregated micro-totals in the SSB.

Where for economic statistics the *System of National Accounts* (SNA) provides a general integrating framework, social statistics lack such an integrating system. Notwithstanding the fact that much effort has been put into developing a coherent set of social indicators, the proposed lists of such indicators cannot be considered as an integrated system in statistical terms. In essence they consist of isolated indicators while the relationships between these indicators remain unquantified and do not have the character of definitional equations (Van Tuinen, 1995).

As the SNA produces authoritative statistical information, it could be promising to develop links between SNA and social statistics, thus facilitating the integration of social statistics by providing anchors. Although in practice economic and social statisticians rarely work together to integrate the two fields of statistics, this does not mean that in theory there are no relevant links between the economic system and social statistics. In particular economic and social statistics meet each other in the common field of supply and demand of services.

Take health, for instance. Social statistics on health describe the actual state of health of the population. This state of health is not only influenced by demographic and socio-economic factors, but also by the consumption of health services. The production of health services is described by the economic statistics in values and volumes. Volume changes are by far the most important data that national accounts produce, as the GDP growth is the measure of the success of economic policy. However, extensive surveying in production unit surveys on quantities is not very popular. That is the main reason why volume changes are estimated by deflating values by price changes. In order to give a complete picture of health in society, the economic information on the volume changes and quantities of the production of health services has to be linked to the state of health, the related demand for health services and medical consumption.

Information on the quantities of the production of services (health services or others) can be obtained from production units or from persons and households. The production of services measured through a production unit survey should be similar to the consumption of services as measured by a household sample survey. Any differences between the two approaches can be quantified and possibly explained. If we are able to explain the difference and quantify each explanation, this will lead to an integrated figure and in turn might result in one figure for one characteristic, and a better quality of results.

After integration of the results obtained from production unit and household surveys on the quantities of the production and consumption of services, we have the opportunity to make use of the already available and commonly accepted integrated data of the SNA to verify the data of production unit and household surveys. As such, we have two related integration processes. The first is the integration of the results on the quantities of the production of services from production unit and household surveys, and the second is the verification of the quantity changes of services by comparing them with the volume changes counted in the SNA. This is not a one-way process; in co-operation with national accounts, we shall develop the estimation of these characteristics further.

The final phase of the integration process consists of the implementation of the established adjustments in the micro-data of production unit and household surveys. The production unit survey will be included in a Statbase for production units and the household survey in the Statbase for persons and households, the SSB. The objective is to make the micro-data consistent with the integrated figures. Consistency between micro and macro-level means that the integrated figure can be reproduced by tabulating the quantities in the micro-data.

In this article, the characteristics of the supply-and-demand matrices are presented, including the actors, roles and transactions. Furthermore, the value of the supply and demand matrices for the policy users is set out. The remainder of the article describes the integration process of quantities of produced and consumed services and the process of the implementation of the adjustments made in the integration process in the micro-data. In the integration process the prototype of the SSB is used to estimate several adjustments. The implementation of the adjustments is restricted to the SSB. The discussion elaborates on the problems of achieving consistency between the volume changes in national accounts and quantity changes in the supply- and-demand matrices.

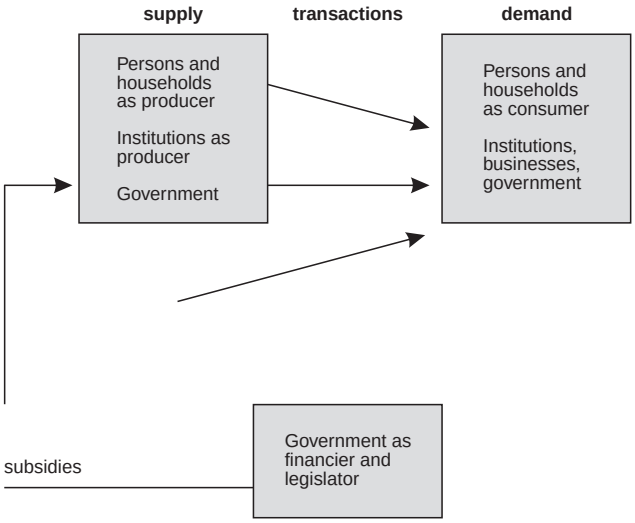
2. The supply and demand matrices

Statistics Netherlands is working on an integrated system of supply and demand matrices of the government and non-profit sector. We are concentrating mainly on NACE sections K, L, M, N and O (legal activities, investigation and security activities, public administration and defence, education, health and social work, other community, social and personal service activities, including recreational, cultural and sporting activities).

Let us take a closer look at the field under study. Firstly, the production units can be characterised mainly as non-market producers. Their activities are mainly financed through government subsidies and contributions by households directly or indirectly via insurance companies, and are strongly steered by government regulations. The statistical units are the local kind of activity units (local KAU), which we shall call institutions. The transactions should be measured in quantities, values (in Dutch guilders, or Euros in the near future) and volume changes.

A second characteristic is the dominant role of the government as an active producer of services. We are aware of a cyclical process, in which the government tends to exchange its role as an active producer of services for a role of financier, and vice versa. This leads to so-called substitution processes between the corresponding activity classes. The statistical units are the government units; together with the local KAU's we shall call them production units. The transactions should be measured in quantities, values and volume changes.

Figure 1
Actors, roles and transactions in the supply-demand matrices for the government and non-profit sector



The third characteristic is the relative importance of voluntary work. Persons and households produce services in three ways: as own-account workers, as voluntary workers and as producers for own consumption. The transactions of own account workers should be measured in the same way as those of production units or government units. The production of services of voluntary workers and the production for own consumption should be measured in quantities and time, as these are below the SNA production boundary and values and volume changes cannot be determined.
Besides being used by persons and households, services are also consumed by production units, institutions and government units. This so-called intermediate consumption should be estimated in the integration process to reduce the differences between production unit and household survey outcomes. The government also plays a role in subsidising persons and households to stimulate the consumption of services; we leave this out of the matrices for the present.

What are the most important variables to be measured for the producers and consumers of the services?

For the legal persons (institutions and government units) who supply services the following variables are relevant: NACE, SNA-sector, institutional sector (non-financial corporations, financial corporation, general government, households, non-profit institutions serving households), type of producer (market producer, producer for own final use, other non-market producer), number and hours of work by professionals, number and hours of work of voluntary workers, region, financial information according to the standards of the *European System of Accounts* 1995 (ESA; Eurostat, 1996) including subsidies and taxes of general or local government, some indication of the production capacity and the quantity and prices of provided services.

For the natural persons (who are engaged in an economic activity in their own right) some information on the quantities of the provided services will be sufficient. A basic classification by age, sex, and income is recommendable.

For the final consumers (persons and households) the following variables are relevant: the quantities of consumed services, a basic classification by age, sex and income, a more precise classification based on the different target groups of the services provided, the satisfaction with the provided services.

For the intermediate consumers (businesses, institutions and government units) the following variables are relevant: the quantities of the consumed services, NACE, SNA-sector, region, financial information according to ESA-standards including subsidies and taxes of general or local government.

3. The value of the supply-demand matrices for the users of statistics

How will users benefit from integrated data on the economic process of the production of services and characteristics of the consumers? In government policy it might support decisions about which regulations and subsidies are needed to achieve policy goals. For instance, decisions on which producers of services should receive subsidies, and how much. And whether the government itself should produce the services or institutions subsidised by the government, or market producers. These choices made by the government have little support from data on the production and consumption of the relevant services. To improve this foundation, the *purposes*, the *production costs* and the *quality* of services and characteristics of the suppliers and consumers will be particularly relevant.

The *purposes* of the services should be measured by the classification of services; services should be classified according to their intended effects on the consumers and their target groups. This is the main reason that we prefer the demand-based concept: we classify services according to the fulfilment of certain needs in society. The only possible reason to look beyond the demand is that differences in the production processes of similar services have different side effects. We will not elaborate on these here.

The *production* costs of the services can be estimated using the financial information on the production units. In this respect we are interested in the production costs of the services and the amount of subsidy from the government. Of course, it will be necessary to be consistent with the results of the national accounts. In practice this will lead to a situation in which the results of national accounts are detailed by supplementary information.

One of the recurring problems in the integration process of production unit and household survey outcomes becomes apparent here: production unit surveys are institution based and the sample is from a business register which only contains information on the primary economic activity of the production unit. Persons and households consume services from production units that produce these services as primary or secondary economic activity.

The *quality* of the produced services is perhaps the most difficult aspect to measure. To measure the quality of services we have to measure to what extent they fulfil their purpose. Several aspects are distinguished in this respect. The first is the need to discover whether the services are consumed by the intended target group. If the majority of the consumers do not belong to the target group, it will not be effective. The other side of the coin is whether the majority of the target group consumes the service. To achieve this, we have to classify the consumers according to characteristics that identify target groups in society. The second aspect is whether the service has the desired effects for those who use it. This can only be measured indirectly, by asking about satisfaction with the provided service.

4. Conflicting results from production unit and household surveys

As mentioned above, the production of services measured through a production unit survey should correspond to some degree with the consumption of services measured by means of a household

sample survey. The following stages of the integration process will be relevant:

– *Harmonisation* which includes the adaptation of definitions, classifications and measure of detail of the information from production unit and household sample surveys;
– *Completion*, which includes corrections for differences in population of the two production units and households covered by the sources;
– *Minimisation of measurement error*, including sampling and non-sampling errors like errors in the sampling framework, observation errors, classification errors, effects of selectivity of response, etc.;
– *Balancing*, i.e. the elimination of remaining (unexplained) discrepancies between the data from the production unit and household sample surveys.

All these steps will give insight into the quality of the available sample survey information and are as such relevant and effective in themselves.

One of the recurring reasons that results differ between production unit and household sample surveys is that the practical definitions of services are not identical. It is quite usual for the production unit surveys to measure the produced services in a highly aggregated way, while in household sample surveys this aggregated concept has no meaning at all, and therefore mostly only a small part of services are measured.

A second explanation is that production unit surveys are institution based and the sample is taken from a business register which only contains information on the units' main economic activities. As persons and households consume services from units that produce these services as primary or secondary economic activity, the resulting estimation of the volume of delivered services will be unequal. To solve this problem production units which produce services as a secondary activity are asked to supply quantity and price information on these services in addition to the relevant information on their main economic activity. In such a way, we combine the relevant information of the supply side and the demand side. The method we have developed for the classification of services indicates which services are produced as a secondary activity in a NACE-class, as we make an inventory of each NACE-class in which services are produced based on various sources.

A third recurring explanation has to do with the national and domestic measurement of the concept. The production unit survey estimates the production of services by domestic production units. This means that services produced for the domestic population of people outside their homeland is not taken into account, while it is included in household sample surveys. The household sample survey only estimates the quantity of services of the domestic population of people, and no estimation of the services consumed by others (e.g. tourists) is taken into account. In short: in the integration process the quantity of the international trade in services should be estimated.

A fourth explanation is that the institutionalised population is omitted from the sample frame of household surveys in the Netherlands, as they are difficult to interview. This category has a large consumption of medical and social work services in particular, and results in large differences between production unit and household sample surveys.

A fifth recurring explanation is that people who die or emigrate in the year under review, have consumed services, but can no longer be interviewed in a household sample survey. Medical consumption in the event of death and consumption of legal services in the event of emigration will be underestimated in household sample surveys if the reference period is longer than a few weeks.

A sixth explanation is the selectivity of non-response in household sample surveys. Non-response is only a problem when it is selective; the magnitude of the non-response is not necessarily a good indicator of selectivity. This becomes a severe problem if the theme of the household sample survey causes the selectivity: for example, if relatively more people with a high medical consumption refuse to respond in a health interview survey. Selective non-response can be identified and corrected for by using information from administrative registers. If registers contain variables with a high correlation to variables in the sample survey and the response is selective on the register variables, these can be used to weight the survey. Exact matching of the administrative registers with the sample survey records produces the best results.

These differences between production unit and household sample surveys are estimated during the integration process and the results of both surveys are corrected accordingly. This leads to one figure for the measured phenomenon. After this integration process, we are able to relate the information on the production of services by production units and the information on the consumption of services by persons and households.

To build the supply-and-demand matrices for services, we started a research programme to explore the possibilities of integrating the statistics on the production and consumption of health services. In the first phase we restricted ourselves to the integration of the quantities of the production and consumption in 1995 of one health service: health care during hospitalisation.

5. An empirical example: bed days in hospital

5.1 Definition of the service, quantification of the difference

We define health care during hospitalisation as:

'A *bed-day*, a period of 24 hours, is the unit for the period, which a patient spends in hospital for cure and/or care, excluding hospitals for mental illnesses. The period starts with the admission and ends with the discharge of the patient. The first bed day is counted as 24 hours after the moment of admission, the second 24 hours later, etc. Healthy babies born during their mothers' hospitalisation are also considered as patients.'

For the estimation of the production of bed-days, we use the 1995 survey on general information on hospitals (SGIH). This integral survey collects information from all official hospitals in the Netherlands. For the estimation of the consumption of bed-days, we use the 1995 health interview survey (HIS), a representative sample survey of the non-institutionalised population (N=9,793). In the SGIH, the number of bed-days is estimated at 15,779,000, while the number of bed days in the HIS is estimated on 11,611,000, a difference of 4,168,000. Below, we shall explain the difference in results and quantify all the explanations separately. Table 1 gives an overview of the adjustment factors.

5.2 Harmonisation

Harmonisation includes the adaptation of definitions, classifications and measure of detail of the information from production unit and household sample surveys.

One of the recurring explanations of the differences between production unit and household sample surveys is that the practical definitions of services are not identical. In the case of bed-days spent in hospital the production unit survey measures the days that are charged to insurance companies. In practice this is always one day more than is counted according to the definition used in this study. As we know the number of hospital admissions, the adjusted number of bed-days can be calculated by diminishing the number of charged days by the number of admissions. This leads to an adjustment of 1,593 thousand bed-days.

The household sample survey asks for the number of nights spent in hospital in the last twelve months, excluding the nights related to

Table 1
The integration of bed-days in hospitals, excluding hospitals for mental illnesses

Bed-days according to production unit survey	15,779,000	11,611,000	Bed days according to household survey
Correction for definition differences			Correction for definition differences
From claim-days to bed-days	–1,593,000	–168	Correction for reference period
		578	Bed-days mother and babies in con-nection with deliveries
Correction for population differences			Correction for population differences
Bed-days non residents	–55	585	Bed-days institutionalised population
		38	Bed-days asylum seekers
		–147	Bed-days of residents in hospitals abroad
Adjustment for measurement errors			Adjustment for measurement errors
		1,258,000	Bed-days of deceased and emigrants
		754	Bed-days of persons in the hospital on the survey date
		–229	Correction for bias due to selective non-response
Remaining difference			Remaining difference
		151	For balancing
	14,131,000	**14,131,000**	

deliveries. Nights spend in hospital is approximately the same definition as the number of bed-days according to the definition. However, as the household sample survey is held during all twelve months of 1995 and the reference period is the previous twelve months, the bed-days counted took place between January 1994 and December 1995. As a result, someone interviewed on 1 January 1995 provided information for the year 1994, while an interview held on 31 December 1995 yielded information on the year 1995. As there is a trend in the number of bed days, the estimation is not entirely correct. The required adjustment was found to number 168,000 bed-days.

Another adjustment is needed for the number of maternity-related bed-days. The household survey did not include bed-days related to deliveries, resulting in the exclusion of a considerable number of bed-days. The *Vademecum of health statistics of the Netherlands* (Statistics Netherlands, 1997) provides information regarding hospital births and corresponding bed-days. In 1995, 62,755 mothers gave birth in hospital; 18,379 of these underwent caesarean section. After a caesarean, a mother spends on average 8.4 days in the hospital. The same number of bed-days was assumed for the babies born by caesarean. For other clinical deliveries the average duration of admission was 3.0 days. Again, the number of bed-days for the corresponding babies was assumed to be the same. The number of bed-days not reported in the household sample survey thus totalled 578 thousand.

5.3 Completion

Completion includes corrections for differences in the target population of both production units and households covered by the sources.

The integration process should include estimation of the quantity of international trade in services. As the production unit survey measures the number of bed-days for the domestic and non-domestic population, we can adjust the figure by subtracting the number of bed-days of the non-domestic population: 55 thousand. The number of bed-days spent in foreign hospitals by the domestic population is estimated in the household sample survey: 147 thousand. This estimation is based on the answers of nine respondents and is therefore not very reliable, but more reliable measurements are lacking.

For practical reasons the institutionalised population is not included in the household sample survey, although they do belong to the target population. Therefore, the net household sample survey had been weighted to the total population (including the institutionalised population), which would imply that the institutionalised population has an average medical consumption. However, this part of the population has a higher medical consumption than average. Fortunately, for the year 1995 a survey had been carried out among institutionalised people aged 65 years and over in homes for the elderly, nursing homes and mental hospitals. The survey covered approximately 70% of the institutionalised population. Our estimation for those 70% is based on this survey. For the other categories, we assume that they have the same medical consumption as the non-institutionalised population according to age and sex. This leads to a adjustment of 585 thousand bed-days.

Another category not covered by the household sample survey are the asylum seekers. There were approximately 42 thousand asylum seekers in the Netherlands in 1995. As no information on their medical consumption is available, we assume that they have the same average number of bed-days as the non-institutionalised population. This leads to an adjustment of 38 thousand bed-days.

5.4 Minimisation of measurement error

Measurement errors include sampling and non-sampling errors like errors in the sampling framework, observation errors, classification errors, effects of selectivity of response, etc.

People who die or emigrate in the year under review, have consumed services, but can no longer be interviewed in a household sample survey. As the reference period of the measurement of the number of bed-days is the last twelve months, this leads to a severe underestimation of the number of bed-days. In particular, it is plausible that people who die have a high medical consumption in the period leading up to death. Assuming that those who died have an average of one admission in hospital and ten bed-days, and assuming that people who emigrate show average medical consumption according to age and sex, the underestimation is approximately 1,258 thousand bed days. Medical consumption in the event of death and consumption of legal services in the event of emigration will be underestimated in household sample surveys.

People who are in hospital when they are asked to participate in the survey will not respond, while they certainly have bed-days. As the number of people who stay in hospital during the year is known, it is possible to estimate the chance that they will miss the survey. As the

survey makes use of proxy interviews, information on the number of bed-days of people in hospital during the survey will be collected. We diminish the number of potential respondents by the number of actual respondents in hospital. Furthermore, we assume that the non-respondents in hospital have the same average number of bed-days as actual respondents in hospital. This results in a adjustment of 754 thousand bed-days.
Another measurement error is the selectivity of non-response in the household sample survey. Selective non-response can be identified and corrected for by using information from administrative registers. The article by Geuzinge, Van Rooijen and Bakker in the present issue elaborates on this subject. They estimate that the selectivity in the household sample survey leads to an overestimation of 419 thousand bed-days. However, as is described in section 6, a correction is needed because of the overlap with other adjustment factors. After a new weighting procedure, the adjustment factor was estimated at 229 thousand bed-days.

6. Consistency between macro and micro data through implementation of established adjustments in the SSB

6.1 Objective of implementation

The final phase of the integration process consisted of the implementation of the established adjustments, as discussed in section 5, in the SSB. The objective was to make the SSB consistent with the integrated figure for bed-days at the macro as well as the micro-level. Consistency at the macro-level means that the integrated figure can be reproduced by tabulating the total number of bed-days from the SSB. Consistency at the micro-level means that the number of bed-days is correct at the record level, i.e. at the level of the individual. This enables us to produce valid cross tabulations of bed-days with background characteristics in the SSB. Consistency at the macro-level was achieved by implementing the adjustments at the micro-level, i.e. the record level. As a consequence, the macro-level and the micro-level were made mutually consistent.

6.2 Methods of implementation

Before the implementation of the established adjustments could be carried out, the health interview survey (HIS) was matched with the SSB. This means that each record from the HIS was matched and subsequently linked with the corresponding record from the SSB. At this point, the uncorrected figure for bed-days, as estimated on the basis of the HIS, had been implemented in the SSB. Subsequently, the established adjustments were carried out at the record-level. This was done in several ways:
– by imputation of bed-days;
– by adjustment of the number of bed-days;
– by re-weighting the survey.

The implementation of the adjustments as well as the employed methods are described below.

Accommodation of non-matched records
A group of 499 records from the HIS had not been matched to the corresponding records in the SSB. In order to prevent loss of information, these records were linked to records of non-respondents that had matched with their corresponding records in the SSB. Matching to records of non-respondents was done on the basis of sex and age, i.e. a non-matched record representing a respondent with a certain combination of sex and age was matched to a record representing a non-respondent with the same combination of sex and age. As a result, the distribution of bed-days according to age and sex was maintained.

Nowadays, Statistics Netherlands draws samples for household surveys directly from the population register, which constitutes the backbone of the SSB. The sample therefore already contains the information that is used in a later phase to link the resulting data with the SSB. As a result, the matching efficiency has increased to almost 100%, thus eliminating to a large extent the aforementioned problem of non-matched records.

Bed-days of residents in foreign hospitals
The HIS included nine respondents who had been admitted to hospital abroad in 1995. The corresponding bed-days had to be removed or isolated as the integrated figure for bed-days referred to domestic hospitals. This adjustment was implemented in the HIS by creating two variables which represented domestic and foreign bed-days respectively. Thus, bed days in foreign hospitals were isolated, enabling separate analyses.

Bed-days of the institutionalised population
In order to adjust for the omission of the institutionalised population, several assumptions were made about the average number of bed-days for the various types of institutions and age categories. The assumptions were partly based on available data sources and could therefore be made operational. Three possibilities were distinguished:
– average number of bed-days equals that of the Dutch population according to age and sex, as measured by the HIS;
– average number of bed-days equals that of the institutionalised population in the survey among elderly in institutions (ITS, 1997) according to age and sex;
– medical treatment is performed within the institution and therefore does not contribute to bed-days in hospital.

The starting point of the adjustment consisted of selecting 356 records from the SSB representing institutionalised persons. The number of records was determined by applying the sample fraction of the gross HIS to the institutionalised population. Thus, this number corresponds to the expected number if the institutionalised population had been part of the sample frame. The records were selected by taking a random sample for each type of institution, the size of which was proportional to the total number of persons in the institution. This method resulted in an accurate representation of the Dutch institutionalised population.

Next, bed days were assigned to the selected persons by means of imputation. In Table 2, the data sources used to impute bed-days are shown for each type of institution. When either the HIS or the survey among elderly in institutions was used to impute bed days, cold deck imputation was used. The bed-days of a record from the SSB representing an institutionalised person were imputed with a record from either the HIS or the survey among elderly in institutions as donors (depending on type of institution and age, see Table 2). The imputation was based on the combination of age and sex.

Table 2
Data sources for imputation of bed-days for the institutionalised population

Type of institution	Data source
Psychiatric institution, older than 65	Elderly in institutions
Nursing home	Elderly in institutions
Home for the elderly	Elderly in institutions
Reception centre for adults, older than 65	Elderly in institutions
Other	Health interview survey

Bed-days of mothers and babies related to deliveries
The HIS did not include maternity-related bed-days, missing 578,000 bed-days as a result. However, other information on deliveries was present in the HIS, enabling imputation of bed-days.

In order to carry out the adjustment, records representing women who had given birth in hospital in the year preceding the survey were selected in the HIS. In addition, the corresponding babies were selected. Subsequently, bed-days had to be imputed in these records according to type of delivery (caesarean or other). Unfortunately, as no information regarding the type of clinical delivery was present in the HIS, no distinction could be made between deliveries by caesarean section and other deliveries. Therefore, deliveries by caesarean were assigned randomly to the selected women and corresponding children. The number of records to which caesareans were assigned was chosen so as to reflect the correct number of caesareans at population level (after weighting of the sample). In these records 8.4 bed days were imputed. Next, the remaining records were designated as women who had given birth in hospital other than by caesarean and their corresponding babies. These records represented approximately the correct number of deliveries at the population level. In these records, 3.0 bed days were imputed.

Adjustment for reference period
In order to adjust for the difference in reference period, 7,800 hospital admissions and 168,000 bed days had to be removed from the HIS. This was done by multiplying the total number of bed-days and the total number of admissions of respondents by a correction factor. Thus, in each record representing a person who had been hospitalised in the year before the survey the number of bed days and admissions was reduced.

Persons in hospital during the survey
As persons who had been hospitalised in 1995 had had a smaller chance to participate in the survey, the ensuing missed observations had to be taken into account. In order make this adjustment, the number of missed observations and the corresponding number of bed-days had to be established. Some of the above-mentioned hospitalised persons had been surveyed indirectly, namely through proxy-interviews. In addition, the average number of hospitalised persons per day was known from the SGIH. The discrepancy between this number and the number represented in the HIS had to be adjusted for. A random sample was drawn from the persons in the HIS who had been hospitalised during the survey (and had been surveyed by means of proxy). Subsequently, the bed-days of these records were imputed in records of non-respondents. Imputation was based on sex and age. This means that bed-days of a record representing a hospitalised person with a certain combination of sex and age were imputed in a record of a non-respondent with the same combination of sex and age.

Deceased and emigrated persons
In 1995, 135,700 people died and 82,878 emigrated. As these people had an average chance of 50% of being approached for the HIS before their death/emigration, many observations had been missed and adjustment was necessary. The SSB contains information on persons withdrawn from the population register in 1995. Therefore, after the HIS had been matched to the SSB, the deceased/emigrated in the HIS could be designated. The discrepancy between the number represented in the HIS and the above-mentioned correct number was adjusted for by a combination of two methods. Firstly, the deceased/emigrated persons in the HIS were given a larger weight by multiplication of the prior weights with a correction factor. Secondly, records of deceased non-respondents were selected and the average number of bed-days of deceased persons (9.9; source SGIH) was imputed. These records were selected according to age and sex so as to correspond to the known population distribution of age and sex of deceased persons.

Asylum seekers
Asylum seekers were not included in the HIS sample frame, but did spend time in domestic hospitals. The required adjustment was carried out by adding a single record to the HIS with the average number of bed days of the Dutch population. Subsequently, a weight was assigned to the record so that it represented the total number of asylum seekers.

New weighting of the HIS
The last adjustment consisted of a new weighting of the HIS to population totals from the SSB. The previous weighting had been carried out in order to adjust for bias due to selective non-response and is described in the article by Geuzinge, Van Rooijen and Bakker in this special issue. The new weighting was necessary for several reasons. Firstly, for practical reasons the existing weights had not been based on population totals from the SSB. Instead, they were based on a selection from the SSB that represented the initial gross sample, that is respondents plus non-respondents. Since the initial gross sample is devoid of errors besides the sampling error, the gross sample had been regarded as a satisfactory weighting frame. However, complete consistency of the weighted HIS with the SSB was desirable, as the HIS had been integrated with the SSB. In addition, the SSB is the most reliable weighting frame as it represents the whole population. Therefore, a calibration of the existing weights was necessary. The adjustment of weights enabled tabulation to SSB totals. Secondly, several adjustments entailed the addition of records to the net sample, for example in the adjustment for the institutionalised population. This enlarged the sample resulting in a reduced average weight. Consequently, adjustments that had been implemented prior to the addition of records, were reduced at the population level, as smaller weights were assigned to the corrected records. Therefore, a variable that indicated the corrected groups of records was included in the weighting model. For example, women who had given birth in hospital by caesarean section were designated by the aforementioned variable. Because of the reduced weights the correct number, i.e. 18,379, was no longer represented by the sample. Subsequently, the sample was weighted to population totals of these groups. In the example this means that, after weighting, the sample again represented 18,379 women who had given birth by caesarean.

It should be noted that the addition of records combined with the new weighting changed the correction for selective non-response. In the article by Geuzinge, Van Rooijen and Bakker (in the present issue), a reduction of 419,000 bed-days is mentioned. However, after the new weighting the reduction amounted to 229,000 bed-days.

7. Conclusion and discussion

One of the advantages of the SSB approach is that a large part of the integration process that has up to now been done at a macro-level, will be performed at a micro-level. However, macro-integration will still be necessary, e.g. for the consistency with national accounts or with the outcomes of production unit surveys. In this article the so-called Supply and demand matrices are described, that is the frame in which the results of production unit and household surveys on the values, prices, volumes and quantities of services are integrated. There are two relevant definitional equations. The first is that the quantities of the production and consumption of services should be similar. In this article we elaborate in particular on the integration process that makes use of this definitional equation. For a large part, the method used for labour accounting systems is adopted to solve the problems of inconsistencies between production unit and household surveys.

The empirical example is restricted to the integration of the quantities of the bed- days in hospital. The general method and an empirical example are presented on the quantities of bed days in hospital. In addition, a method is presented to achieve consistency between macro-totals and the micro-data of the quantities of the consumed services. In the integration process several adjustment factors are estimated. After the determination of the integrated quantities of services, it is necessary to attain consistency between the macro-totals and the aggregated micro-totals in the SSB.

The final phase of the integration process consisted of the implementation of the established adjustments, as discussed in section 5, in the SSB. The objective was to make the SSB consistent with the integrated number of bed days at the macro-level as well as at the micro-level. Consistency at the macro-level means that the integrated figure can be reproduced by tabulating the total number of bed days from the SSB. Consistency at the micro-level means that the number of bed days is correct at the record level, i.e. at the level of the individual. This enables the production of valid cross tabulation of bed days with background characteristics in the SSB. Consistency at the macro-level was achieved by implementation of the adjustments at the micro-level, i.e. the record-level. As a consequence, the macro-level and the micro-level were made mutually consistent.

Several adjustments, such as imputation of bed days, were carried out while taking into account the relationship between bed days and the background variables sex and age. However, the relationship with several other background variables, such as ethnicity, was not taken into consideration. Therefore, it might be interesting to establish the extent to which the adjustments have affected these relationships. At the moment, the effect of the described adjustments on the relationship between bed days and background variables is being evaluated.

The second relevant definitional equation is value = volume * price, and related to this the equation value index = volume index * price index. This equation is used to produce measures of volumes at constant prices. The resulting volume index gives a reliable estimate of the changes in GDP . In order to keep up consistency with the national accounts, the changes in quantities of the produced and consumed services should be aggregated after weighting to the volume changes of the national accounts based on deflated values.

Most data underlying the national accounts are in current prices and are transformed into volume information by deflating them with price indices. This procedure implies the assumption that prices are measured representatively: a statistically correct sample of enterprises or persons has been surveyed to obtain information on the prices of a set of services that undergo the same price changes as the product groups in the national accounts. In many cases such reliable price data are absent, in particular in the field of services of the government and non-profit organisations. In cases where the measurement of prices is inexact, these prices have an enormous impact on the volume estimates of the national accounts. Actual quantity information on the produced and consumed services should be balanced with the data of national accounts (De Boer, Van Nunspeet & Takema, 1999). Many data problems still need to be solved, in particular the problem of defining similar homogeneous products and taking into account the quality changes in the products. In collaboration with national accounts, a method will be developed to achieve more consistency.

References

De Boer, S., W. Van Nunspeet & T. Takema, 1999, *Supply and use tables in current and constant prices for the Netherlands: an experience of fifteen years* (Voorburg/Heerlen: Statistics Netherlands).

Eurostat, 1996, *European System of Accounts 1995* (Luxembourg: Eurostat).

ITS, 1997, Veldwerkverslag project '*Onderzoek naar ouderen in instellingen* 1996' *(Fieldwork report project'Survey on Elderly in Institutions 1996')* (Nijmegen: ITS).

SNA, 1993, *System of National Accounts 1993* (Brussels/Luxembourg, New York, Paris, Washington D.C.: CEC, IMF, OECD, UN & WB).

Statistics Netherlands, 1997, *Vademecum of health statistics of the Netherlands 1997* ('s-Gravenhage: SDU).

Van Tuinen, H.K., Social indicators, social surveys and integration of social statistics, In: *Netherlands Official Statistics*, vol. 10, autumn 1995, pp.5–22

Data security, privacy and the SSB

Pieter G. Al and Jan Willem Altena

1. Introduction

Because ultimately the SSB will comprise a very detailed picture of every inhabitant of the Netherlands, data security and privacy are very important issues. Statistics Netherlands cannot risk individual data being disclosed as the support of the Dutch and their government will then be withdrawn. But besides this threatened loss of support, there are legal conditions which prevent Statistics Netherlands from publishing individual data; these are described in section 2. Section 3 is devoted to informed consent and we finish with a section in which the security measures are described in detail.

2. Legal preconditions

In the seventies there was a growing concern in the Netherlands about the protection of privacy. Although the 1971 general census prompted public debate on the subject, this affected the response to that census only slightly: The final non-response rate amounted to 0.026 per cent. However, the fast growing concern caused a postponement of the 1981 census because of the risk of a non-response rate as high as 26 per cent of the population, and led to the decision to abandon the censuses in later decades.

The public debate itself started a process of legislation on the subject of protection of privacy. Today the protection of privacy is well regulated in the Netherlands. The Act on Personal Data Registrations (WPR) regulating the maintenance and use of registrations of personal data, was adopted in 1988. The so-called Registration Chamber supervises the enforcement of the act. In 2000 the WPR will be replaced by the Personal Data Protection Act (WBP) [1)].

Besides the general legislation on privacy protection, some specific rules are laid down in the legislation concerning the Central Commission for Statistics, the Central Bureau of Statistics and the production of statistics. Furthermore Statistics Netherlands endorses the Declaration on Professional Ethics of the International Statistical Institute (ISI) published in 1986 (International Statistical Review, 1986).

The main regulations in these Acts and the Declaration about the use of individual data and the protection of privacy are summed up below.

2.1 CBS Act

The very existence of the Central Bureau of Statistics (or Statistics Netherlands as it is now called) is enforced by a special act. Two articles are relevant in the current context. Section 9 (see box) gives Statistics Netherlands the right to use data from government institutions and the right to use the social security number. Such data from government institutions are, for example, fiscal information and information regarding social security schemes. Statistics Netherlands receives much of this information on an individual basis, thus making it possible to link these data with other information within the Bureau. One very efficient way for linking these data is matching with the aid of the social-security number. As this number may only be used under specified legislation, it cannot be used as a general number for identification, as is for example the case in the United States. All users of the social security number may only use the number for their own purpose.

Some Sections from the CBS Act

Section 9

1. The CBS is authorised to use, for statistical purposes, data from the records of State institutions and agencies. The CBS shall exercise this authority in accordance with Our Minister under whose responsibility the records in question fall.
2. The CBS shall be entitled to include the social security and fiscal number, as referred to in Section 47b(3) of the Law on State Tax, in personal data files and use it for statistical purposes. The CBS shall be entitled to use this social security and fiscal number in contacts with persons and official bodies to the extent that they themselves are authorised to use the number in a personal data file.

Section 11

1. The data received by the CBS in connection with the performance of its responsibilities shall be used solely for statistical purposes.
2. The data referred to in subsection 1 shall not be furnished to other persons than those charged with carrying out the responsibilities of the CBS.
3. …

Section 11 sets rules for data security at Statistics Netherlands. The section governs data security from Statistics Netherlands to the outside world, and is an important section in respect of data dissemination (see Statistics Netherlands, 1999 for more details).

2.2 The Act on Personal Data Registrations (WPR)

The act on Personal Data Registrations (WPR) governs data security inside Statistics Netherlands. See the box below for the relevant sections of this law.

Section 8 applies to the data security from Statistics Netherlands to the outside world and is a kind of a duplication of Section 11 of the CBS Act. Section 6.2 is relevant for the production process within Statistics Netherlands: everyone within Statistics Netherlands

Some sections of the Act on Personal Data Registrations (WPR)

Section 6. 2.

Personal data files which are held by the organisation will be made available only to those persons who are authorised to receive such data in order to execute their task.

Section 8.

The organisation which holds personal data files will procure the necessary technical and organisational provisions against loss or interference with these files and against unauthorised data access, data alterations and data dissemination. Staff members must take accordingly protective measures for all or part of the data processing apparatus for which they are responsible.

should have access only to the data which they need to do their job. This complicates the production process. For instance at the start of the statistical process often the identifying data (name, address, and social security number) and statistical data are combined in one file. Section 6.2 states that if this combination is no longer necessary, the identifying variables should be removed. This is often the case in later stages of the statistical process.

The proposed Data Protection Act (WBP) has the same kind of sections, so the measures for data protection will be the same.

2.3 ISI declaration on professional ethics

Besides the Dutch legislation on privacy protection and the legislation on official statistics in the Netherlands, there are several European rules on statistics and the UN principles on official statistics. On the subject of information for the public, the declaration on professional ethics of the International Statistical Institute is one of the clearest statements. Although the rules of this declaration cannot be considered as a legal obligation, Statistics Netherlands considers them as a valuable code of behaviour. With respect to the combination of data from various sources, the most important rules are those in chapter 4 of the declaration, concerning the obligations to subjects of statistical surveys (see box). These prescribe proper information to respondents to sample surveys about the use of the information they provide.

Some articles of the ISI declaration on professional ethics

Section 4.1 Avoiding undue intrusion.

Statisticians should be aware of the intrusive potential of some or their work. They have no special entitlement to study all phenomena. The advantage of knowledge and the pursuit of information are not themselves sufficient justifications for overriding other social and cultural values.

Section 4.2 Obtaining informed consent.

Statistical inquiries involving the active participation of human subjects should be based as far as possible on their freely given informed consent. Even if participation is required by law, it should still be as informed as possible. In voluntary inquiries, subjects should not be under the impression that they are required to participate; they should be aware of their entitlement to refuse at any stage for whatever reason and to withdraw data just supplied. Information that would be likely to affect a subject's willingness to participate should not be deliberately withheld.

3. Informed consent

Under the CBS Act Statistics Netherlands is entitled to use, for statistical purposes, data from the records of state institutions and agencies. Statistics Netherlands is also entitled to micro-link the data from different sources using social security and fiscal numbers. Although in general the WPR restricts the use of personal data to the original purpose of the data collection, an exemption is made for the processing of those data for historical, statistical and scientific purposes. In the same way an exemptions is made for keeping the data for a longer period if this serves historical, statistical and scientific purposes. Based on this legal feasibility, in the past decade Statistics Netherlands has done some research with linked data from different sources, including data from sample surveys among households.

The endorsement by Statistics Netherlands of the ISI declaration on professional ethics is highly important for the use of data from household sample surveys. The rules of this declaration prescribe that the respondent shall be informed about the use that will be made of the data. As long as these data are used to produce so-called stovepipe statistics, the respondent may be assumed to be informed about the use of the data. At the beginning of the interview respondents are always told about the objectives of the survey they have been invited for.

In 1996 some first experiments with a broad linking of data from various sources were carried out to investigate the possibilities of a micro database like the SSB. Data from sample surveys among household were used without any further announcement to the respondents. When these experiments appeared to be successful, a plan was made to develop a new integral statistical system based on micro-data in the form of the SSB. In 1998, a procedure was developed to inform all respondents about the wider use that will be made of the data they provide.

This procedure consists of two elements. The first is a clear statement in the introductory letter inviting people to participate in the survey. The second is a brochure with a much more elaborate description of the use of individual data by Statistics Netherlands. The brochure explains clearly how confidentiality is guaranteed.

In the introductory letter, addressees are told that Statistics Netherlands may link the provided data to other data available and that Statistics Netherlands safeguards the privacy of each respondent. The passage on the matching of data reads as follows:

> Statistics Netherlands not only collects data but also receives data files from other institutions, e.g. from the population register, employment agencies, social security agencies and public and private sector payroll administrations. Information from sample surveys among households is linked to register based information. This integration of data from various sources enables Statistics Netherlands to assemble efficiently reliable social and economic statistics which provide an adequate quantitative representation of Dutch society.

The twelve-page brochure is sent to respondents on request or handed out by the interviewer. It contains information on the purposes of CBS surveys and how the results are published without any disclosure of individual data. It also contains a list of other data sources used by Statistics Netherlands and a clear statement about the linking of data from different sources. Lastly, the brochure gives a description of the measures for data security and privacy enhancement.

This procedure has been used for all sample surveys since mid 1999, fulfilling the requirement of the ISI declaration rules on informed consent. A study into the effects of this procedure on the response rate of the sample surveys indicated a slight decline of response rates. However, it is not completely clear what caused this as at the same time other things were changed in the introductory letters and the survey processes.

4. Security measures

To achieve a sufficient level of data security, measures are taken on the following points:

Staff

Since 1 January 1998, the Director-General of Statistics Netherlands is obliged to swear in every new member of staff at Statistics Netherlands. The oath consists of the following text:

> I swear/promise that I will execute the tasks appointed to me dutifully and conscientiously, and that I will protect the secrecy of confidential matters of which I gain knowledge from all persons except those to whom I have to inform *ex officio.*

This oath is intended to ensure that every member of the staff realises that data-security is an important issue. Members of the staff responsible for data security are also obliged to attend a course on this issue. One of the aims of this course is to make the staff aware of the pitfalls of data disclosure. The course is targeted at data security from Statistics Netherlands to the outside world. See also Statistics Netherlands, 1999.

Computer network
The first measure to ensure the data protection from the outside world is a strict check of people entering the buildings. Each member of the staff has an identity card enabling him/her to enter the building. Visitors have to check in at reception and identify themselves with an official document.

A second measure is that the computer network is not connected to the outside world. Therefore, it is not possible to enter the internal network of Statistics Netherlands from outside. Of course, Statistics Netherlands also has an external network (among others for its website), but this network has no - real time - connection with the inner network. This measure, and the third measure of limitation of e-mail facilities, ensures that it is highly improbable that unsecured data can leave Statistics Netherlands.

The third measure of limitation of e-mail facilities consists of not having the right to send e-mails with attachments. This limitation is only for staff who have access to data with identifying variables. With this limitation, it is highly improbable that individual data can be exported by e-mail. Moreover, if a member of the staff sends e-mail with individual data in plain text out of Statistics Netherlands, criminal intent is easily proven.

The fourth measure is restricting access rights on the internal network. These rights are restricted in place and time. Only staff working directly on the SSB have access to the part of the network on which the SSB is produced, and they only have access to the data they need. The staff who match data from different sources only have access to the data on which the matches are made and not to the statistical data on the same subjects. Conversely, staff who for example perform statistical analysis on the linked data, only have access to the statistical variables of the linked cases, and not to the identifying variables.

The data
We shall discuss several methods for safeguarding the data themselves below, concluding with the method we use for the SSB. The objective is to separate the identifying variables from the statistical variables, this is one of the consequences of Section 6.2 of the Act on Personal Data Registrations (see section). We subsequently describe how record linking is carried out in a split-data environment. Before we can use encryption for the data security, we need a system for key management. At this moment, we do not have a detailed procedure for this, but we hope to develop such a system in the coming months.

Methods of splitting the data

A naive method.
A naive method is to strip the identifying variables from the file. Symbolically:

$[I,S] \rightarrow [S]$

With this method, identifying is no longer possible, but it is also impossible to link the data with other data. This is not a desired situation, as one of the essential properties of the SSB is precisely the possibility of micro-matching the data from different sources.

An often used method
A second method is to split the file into two parts: one with the identifying variables and one with the statistical ones. The two parts of a record are related by a number: The set of identifying variables I of record A is augmented with a number N_A and the set S of statistical variables of record A is also augmented with this number. The number N_A should be unrelated with the identifying variables; a safe choice is a unique random number. Symbolically:

$[I,S] \rightarrow [N,I,S] \rightarrow [N,I],[N,S]$

This method is very simple and often used, but is not satisfactory for the SSB as disclosure is very easy if the person who has access to the identifying variables conspires with someone who has access to the statistical variables. In addition, staff who are responsible for maintaining the computer network have access to the data. The next method solves these problems:

A method which uses encryption
A third method is splitting the file in two parts: one with the identifying variables and one with the statistical ones. The two parts of a record are related by a two numbers: The set of identifying variables I of record A is augmented with a number N_A and the set S of statistical variables of record A is augmented with an encryption $E(N_A)$ of this number. The encryption is done with one key. Symbolically:

$[I,S] \rightarrow [N,I,S,] \rightarrow [N,E(N),I,S] \rightarrow [N,I],[E(N),S]$

This method has the advantage that without knowledge of the key, it is impossible to relate the identifying and the statistical variables. This is – of course – only true when the encryption method is strong enough. A disadvantage is that if the encryption method is broken, the whole procedure of splitting the data must be renewed.

The first SSB method
The method we used with the first prototype of the SSB was an extension of the second method. Instead of one number, we used two –unrelated – numbers. The splitting of the file results in three parts instead of two:
– One part with the identifying variables and for each record a number N_I
– One part with the statistical variables and for each record a number N_S
– And a table with the numbers N_I and N_S for each record (the relation table)

Symbolically:

$[I,S] \rightarrow [N_I,I,N_S,S] \rightarrow [N_I,I],[N_S,S],[N_I,N_S]$

This method has the advantage that three people are needed for a conspiracy (one with access to the identifying variables, etc.); but the disadvantage is that the staff responsible for running the computer network have access to the data.
The next method solves this last problem:

The ultimate SSB method
The method we currently use for the SSB is an extension of the former method. The only addition is that the table with the numbers N_I and N_S is encrypted, symbolically:

$[I,S] \rightarrow [N_I,I,N_S,S] \rightarrow [N_I,I],[N_S,S],[N_I,N_S] \rightarrow [N_I,I],[N_S,S],E([N_I,N_S])$

The encryption of the relation table is done on the whole file at once and not row by row. If after a period the method of encryption is broken, it is very easy to use a stronger method. To use this method only a decryption and a new encryption is needed. There is no need for bring the identifying and statistical variables together.

Record matching in a data-split environment
The method we use for the SSB can also be used for any ordinary survey. Within the context of the SSB, record linking is essential so the method is extended to cover record linkage activities. The first step in linking of two surveys is to match the two parts with the

identifying variables from survey A and survey B respectively, naturally on the basis of the identifying variables. The set of these variables differs from source to source. The result is a table with the two record numbers and two files with records that could not be linked. Symbolically:

$[N_{IA},I_A]$ $[N_{IB},I_B]$ $[N_{IA},N_{IB}],[N_{IA},\ ],[N_{IS},\ ]$

Where the tensor-symbol ( ) represents the process of matching the surveys A and B on their identifying variables. The table $[N_{IA},\ ]$ represents the records from survey A which could not be linked to records form survey B, etc. In the SSB we call these last two tables the dregs [2].

The next step is linking the statistical variables of the two sources: first, the relation table with the two columns N_I and N_S of survey A is decrypted. This is also done for survey B. Thereafter the two identifiers of the statistical variables of the two surveys are linked with the aid of the relation table of the identifiers of the identifying variables of the surveys A and B. Thereafter the statistical variables can be linked. Symbolically:

1. $[N_{IA},N_{SA}],[N_{IB},N_{SB}],[N_{IA},N_{IB}]$ $[N_{SA},N_{SB}]$
2. $[N_{SA},N_{SB}],[N_{SA},S_A],[N_{SB},S_B]$ $[S_A,S_B]$

Notes

1) The parliament already accepted the act. In 2000 the Senate will vote on the WBP.
2) During this phase of the SSB we only perform exact matches. In a latter phase (final integration) we will also apply non-exact matching procedures. In that phase we will use the dregs.

References

International Statistical Review, volume 54, 2, pp. 227-242, 1986 (Voorburg).

Statistics Netherlands, *Netherlands Official Statistics*, volume 14, spring 1999 (Voorburg/Heerlen).

Appendix
Sources for micro-data files of persons, families, households, jobs, benefits and living quarters

	Source	Statistical unit	Population	Integral or sample	Reference period
1.	Population register (municipal basic administration of personal data) (PR)	Person	Population of the Netherlands	Integral	1 January
2.	Vital and migration events (municipal basic administration of personal data) (VME)	Person	Population of the Netherlands	Integral	Annual file
3.	Administration of employee insurance schemes – jobs and payrolls (AEIS-jobs; AEIS-payrolls)	Job	Jobs of private sector employees	Integral	Annual jobs file and annual payroll file
4.	Administration of employee insurance schemes – unemployment benefits (AEIS-UB)	Benefit	Unemployment benefits private sector	Integral	Annual file
5.	Administration of employee insurance schemes – disablement benefits (AEIS-DB)	Benefit	Disablement benefits private sector	Integral	Annual file
6.	Administration of public sector employees disablement benefits (APEDB)	Benefit	Disablement benefits public sector	Integral	Annual file
7.	Social assistance benefits administration (SABA)	Benefit	Social assistance benefits	Integral	Annual file
8.	Employment agency files (EAF)	Person	Persons registered at employment agencies	Integral	Annual file
9.	Students at vocational colleges and universities (SVCU)	Person	Students enrolled in vocational and university education	Integral	Annual file
10.	Income information system (IIS)	Person	Population of the Netherlands	Sample (1%)	36 891
11.	Register of addresses of institutional households (AIH)	Address	Institutions and homes in the Netherlands	Integral	36 891
12.	Annual survey on employment and earnings (ASEE)	Job	Jobs of employees	Selective sample (50%)	Annual file
13.	Labour force survey (LFS)	Person	Population in private households aged 15 years or over	Sample (1%)	Annual file
14.	Household budget survey (HBS)	Household	Private households	Sample (0.03%)	Annual file
15.	Socio-economic panel survey (SEPS)	Person	Population in private households aged 15 years or over	Sample (0.03%)	Annual file
16.	Continuous quality of life survey (CQLS)	Person	Population in private households	Sample (1%)	Annual file
17.	Housing register (HR)	Living quarters	Dwelling stock of the Netherlands	Integral	1 January
18.	Valuation of real estate registration system (VRERS)	Real estate	Real estate of the Netherlands	Integral	1 January
19.	Geographic base file (GBF)	Address	Geographic codes of addresses in the Netherlands	Integral	1st January