Austrian Journal of Statistics
Volume 33 (2004), Number 1&2, 69-88

The Dutch Virtual Census 2001: A new approach by combining Administrative Registers and Household Sample Surveys[1]

Frank Linder
Statistics Netherlands, Voorburg

Abstract: The last traditional population census in the Netherlands dates from 1971. Since then the willingness of citizens to participate in a census started to decline because of privacy considerations. Census information is still necessary for policy and research purposes. An alternative for the Census Programme 2001 was found in the Virtual Census. The advantages are its low response burden on the population and considerably lower costs.
The Virtual Census 2001 uses the Social Statistical Database (SSD) as its source. The SSD contains a huge amount of data on demographic and socio-economic issues. It is constructed by micro-linking several administrative registers and household sample surveys. A micro-integration process ensures coherence, consistency and completeness of the SSD-data. With a new method of repeated weighting, consistency can be achieved between register counts and sample survey estimates on an aggregated level.

Zusammenfassung: In den Niederlanden wurde zuletzt im Jahre 1971 eine traditonelle Volkszählung durchgeführt. Seither hat die Bereitschaft der Bevölkerung, an einer neuenVolkszählung teilzunehmen aus Gründen des Schutzes der Privatatmosphere gründlich abgenommen. Volkszählungs-ergebnisse bleiben aber wichtig für Politik und Wissenschaft. Deshalb werden die Daten für das Volkszählungsprogramm 2001 mittels eines Virtuellen Zensus aufbereitet. Dieser Zensus hat gegenüber einer traditionellen Volkszählung den Vorzug, die Ergebnisse mit einer geringen Belastung der Bevölkerung und mit weit geringeren Aufbereitungskosten zur Verfügung zu stellen.
Wichtige Informationsquelle für den Virtuellen Zensus ist die sozial-statistische Datei (SSD). Sie enthält eine Vielfalt an demografischen und sozial-ökonomischen Daten, die mittels Verknüpfung von Einzeldaten („micro-linking“) aus verschiedenen Register- und Stichprobenerhebungen hergestellt sind. Das Verwenden von micro-linking stellt Kohärenz, Konsistenz und Vollständigkeit der Daten in dieser Datei sicher. Mittels einer neu entwickelten Methode des wiederholten Gewichtens wird

[1] This article is an abridged version of the paper ‘The Dutch Virtual Census 2001, a new approach by combining Administrative Registers and Household Sample Surveys’ (Linder, 2003), presented at the DIECOFIS Workshop on Data Integration and Record Matching, Vienna, Austria, 13-14 November 2003.

Konsistenz zwischen Registerzählungen und den hochgerechneten Daten aus Stichprobenschätzungen erreicht.

Keywords: Administrative Registers, Consistent Table Estimates, Household Sample Surveys, Micro-integration, Micro-linking, Netherlands, Population Censuses, Repeated Weighting, Social Statistical Database.

1 Introduction

Until the nineteen seventies, population censuses in the Netherlands were organized by *national field enumeration*. Then a growing distrust in the objectives of a government collecting all sorts of information about its citizens came about. It marked an era in which society became less co-operative, forcing Statistics Netherlands to find alternatives for the traditional population census. Instead of field enumeration, Statistics Netherlands explored *administrative registers* and *sample surveys* as new data sources in order to fulfil the need of census information.

For the 1981 and 1991 Census Rounds demographic data were drawn from the Population Register. Data on socio-economic characteristics, such as on labour and education, were provided by the Labour Force Survey. These sources, however, were used separately, which means that no special attention was paid to coherence of the information at the micro-level. Moreover, table totals in one source could be different from corresponding totals in the other. To overcome this consistency problem, table results were reweighted to the level of the Population Register totals.

For the Census 2001 Programme (Eurostat, 1999) Statistics Netherlands launched a *new approach*, which is unique in Europe (Eurostat, 2003). One of the most important achievements of the nineties in the area of social statistics is that an increasing amount of socio-economic statistical information can be acquired from administrative registers. Comprehensive and detailed information is now available on employment and social security. By *micro-linkage* and *micro-integration* of demographic and socio-economic data from a wide variety of administrative registers and sample surveys Statistics Netherlands created a so-called *Social Statistical Database* (SSD). The SSD contains coherent and detailed information on persons, households, jobs and (social) benefits. Therefore it is an appropriate data source for the Population Census of 2001 (Vliegen and Van der Laan, 1999). Consistency of sample survey and register (sub) totals is ensured by a newly developed method of *repeated weighting*.

The next section provides a short historical perspective of the Dutch Census. Section 3 describes the underlying data sources of the SSD. Section 4 goes into the process of micro-linkage and micro-integration of these data sources. Section 5 discusses the creation of the SSD. Section 6 deals with methods used to compile a census based on the SSD. The following subjects are discussed: SSD-harmonisation with the census guidelines, register-counting in the case of register variables, and the repeated weighting method where survey variables are involved. Section 7 comes with some concluding remarks. Appendix 1 gives an overview of the data sources used for the Census 2001.

2 A short historical perspective of the Dutch Census

The first *traditional Population Census* in the Netherlands, according to the Royal Decree of 1828, was held in 1829 on the basis of national field enumeration. From then on, censuses were organized once every ten years. There were six more censuses under the responsibility of the Ministry of the Interior in 1839, 1849, 1859, 1869, 1879 and 1889. Shortly after its establishment in 1899 Statistics Netherlands was put in charge of the organization of the Census. Six more traditional censuses were carried out in the 20th century: in 1909, 1920, 1930, and after a break in wartime, in 1947, 1960, and the last one in 1971.

Traditional censuses were needed for two reasons. First, to determine the size of the population and to get statistical information about its socio-economic characteristics. Second, to check and update the municipal population registers.

With the 1971 Census it turned out that civilians were less willing to participate, mainly in the cities. Together there were almost 300 thousand non-participants, which is 2.3 percent of the population. The reason was a growing distrust of the government and a fear that census responses would negatively affect their civil rights, social benefits or privileges. Subsequently more and more people became convinced that a census was a serious invasion of their privacy. This led to the cancellation of the 1981 Census, because pilot studies indicated that the non-response rate would average 25 percent or worse. Finally, in 1991 the Dutch Parliament decided to withdraw the Census Act. As a result Statistics Netherlands was relieved of its legal obligation to carry out a traditional population census. Another decisive motive for abandoning the traditional census, apart from the privacy aspects and the increasing non-response, was the fact that alternative data sources became available which were much cheaper. A conventional census nowadays would cost about 300 million euros, while using the available data sources and methods cost only a fraction of it: 3 million euros.

For the 1981 and 1991 Census a substitutive Census Programme was implemented (Vliegen and Van de Stadt, 1988, and Corbey, 1994). Demographic information was extracted from the Population Register and socio-economic information from the Labour Force Survey. The survey data were reweighted to the level of the Population Register totals. Whereas in the past conventional censuses were used to determine population size and to update population registers, in the period after the 1971 Census the Population Register is used to set the new benchmark for the population size.

For the 2001 Census Programme Statistics Netherlands has *innovated* its *data collection* and *data processing* methods. This is discussed in the sections 4 to 6.

An extensive discourse of the history of Dutch Censuses is to be found in Statistics Netherlands (2002).

3 Data sources 2001

The Virtual Census 2001 is based on the Social Statistical Database (SSD). The underlying data sources of the SSD are described in the following subsections. An overview of these data sources is presented in Appendix 1, table 3.

3.1 Population Register (PR)

The *Population Register* (PR) contains demographic information on every inhabitant of the Netherlands (Prins, 2000). The PR is built from the *municipal population registers*, which are of outstanding quality nowadays. Municipalities have a major incentive to record all their inhabitants because the allocation of central government funds is mainly based on population size.

The Census Programme requires a population concept in which people are counted at the place where they usually reside. This is the *de jure* concept. Persons who are temporarily absent at the time of the census, migrants, homeless people and others who have no 'usual' place of residence are also to be included in the *de jure* population. In the latter case, according to the guidelines of the Census Programme (Eurostat, 1999), an address ought to be assigned where they are enumerated or registered.

The PR meets the requirements of the *de jure* concept because it registers the population at the usual place of residence. Moreover, the PR also encompasses (nearly) all homeless people. They are sometimes registered at their shelter address, for example the Salvation Army, or at the location from where they receive social assistance benefits. These are points where the PR compares favourably to a traditional census. With field enumeration, there is always a chance that the homeless people are overlooked or that persons, who are temporarily staying at the enumeration address but who usually reside elsewhere, are counted by mistake.

Even though the PR seeks to optimally record every person in the population, it is by no means perfect. People may move to live elsewhere and forget to notify the authorities. Therefore, municipal population registers are not always up-to-date. Another example of improper registration in the PR is that of two persons who are registered at separate addresses, but who actually live together. If one person is employed and the other is on welfare they have a financial incentive to be registered at different addresses. This is because the person receiving benefits might loose them when the social security agency finds out they are living together.

An important population group that the PR misses are the people who live in the country without the authorities' knowledge, many staying illegally. The non-registered population group is not present in the census population. *Illegal residents* pose a problem for statistical offices because, on the one hand, they participate in the economy and as such they are included in economic statistics. On the other hand, they are not covered by demographic statistics. It is very unlikely that they would be enumerated in a traditional census though. Statistics Netherlands has made an attempt to estimate the size of the illegal population, but since there is hardly any information this proved very difficult. The official estimate of the number of illegal residents on 1 January 2001 by Statistics Netherlands is one with a wide margin: between 46 thousand and 116 thousand people (Hoogteijling, 2002).

The PR also provides *household information*, such as household size, household composition, household type and household status. The household type indicates whether someone is living in a private or in an institutional household. The household status refers to the position (e.g. child, spouse, cohabitant, single parent and living alone) of a person living in a private household. For about 93 percent of all households it is easy to determine the household composition and status. For the remaining households the relationship of the persons within the household is not quite clear. For

example, it is difficult to distinguish between partners living together and two students sharing an apartment. In these cases the household variables are imputed by means of a probabilistic model that is based on known relationships between persons in comparable households in the Labour Force Survey (Harmsen and Israëls, 2003).

3.2 Jobs register and sample survey (employees)

Until the Census of 1971, data on economic activity of the population were always collected by enumeration. For the two censuses afterwards, in 1981 and 1991, the Labour Force Survey (sampling fraction 5 percent in 1981, and 1 percent in 1991) served as a source for this kind of information.

In the nineties Statistics Netherlands got a complete *register (volume based)* on jobs at its disposal, the *Employee Insurance Schemes Registration System for Employees (EIS-Employees)*. The use of a register puts an end to the problems with sampling errors, but most administrations are not designed for statistical purposes. Therefore some data pollution is unavoidable. This means that extensive checks and edits are required before the data are fit for statistical use.

One drawback of the EIS-employees register as source for the census is that it lacks information on two variables which are needed for the Census Programme, namely 'time usually worked' and 'place of work'. These variables can be found in the *Survey on Employment and Earnings (SEE)*. The SEE is a large-scale survey among enterprises, in which the data are mainly obtained by electronic data interchange (EDI) from payroll administrations. The survey contains information about earnings and working hours of employees as well as some characteristics of their jobs. The SEE has a complicated sampling design: for most of the large enterprises the data are available on a register basis, whereas a sample is taken for the smaller enterprises.

3.3 Jobs register (self-employed persons)

Information on the jobs of self-employed persons (with or without personnel) is to be found in the *register of final income tax assessments on profits of self-employed persons (FiTap)*. Unfortunately, the FiTap-register does not possess data on the exact period of income. Therefore, it is assumed that those who were registered somewhere in 2000 were also self-employed persons on the Census reference date of 1 January 2001. But the assumption may lead to an overestimation of the number self-employed persons on this date. While compiling the Census 2001 table programme, information was still lacking for approximately 40 thousand self-employed persons (5%). Their tax assessment was not yet ready, probably because of disputes with the fiscal authorities.

3.4 FiBase-register

The *FiBase-register* is a *fiscal administration*. It stores data on labour and social security income that is subject to advance tax payments. The register also covers life insurances and pensions from former activities. It is therefore appropriate to trace those

persons in the economically inactive population who are retired, which is information the Census asks for. The FiBase-register is also used to complete missing data on jobs and benefits in the micro-integration process (see subsection 4.2).

3.5 Labour Force Survey (LFS)

The *Labour Force Survey (LFS)* is a *household sample survey*. It is needed for census information that is not (yet) available in registers. It concerns census variables such as occupation and educational attainment levels. The LFS is also used to define that part of the economically active population that is unemployed, and to define those in the economically inactive population who are engaged in family duties as their main current activity or who attend educational institutions full-time.

The LFS is a survey on private households, in which the survey population is restricted to persons aged 15 years and older. It is a continuous survey, meaning that sampling and surveying of persons is spread throughout the year. The sample size is actually quite small; some 100 thousand persons are sampled, which is approximately one percent of the total population. The consequence is that estimation for small subpopulations at a detailed level, which is often asked for in the Census Table Programme 2001, may be unreliable or even impossible. For this reason two LFS-surveys, 2000 and 2001, were joined to create more mass. In fact, information up to one year before the census reference date (1 January 2001) and up to one year after the reference date has been gathered in this way. In general the above mentioned LFS census variables are relatively stable within the period of a year, so that it can be assumed that they represent the situation at reference date without much error.

3.6 Remaining data sources

Some registers only play an indirect part in the 2001 Census. It concerns the Employee Insurance Schemes Registration System–Disablement insurance (EIS-DI), Employee Insurance Schemes Registration System–Unemployment insurance (EIS-UI) and the Social Assistance Benefits Administration (SABA). No variables of these sources are used for the census, but the registers are essential in the micro-integration process of jobs and benefits. This will be discussed in detail in subsection 4.2.

4 Combining data sources, micro-linkage and micro-integration

Policymakers and scientists nowadays increasingly demand *comprehensive* and *coherent* statistical information that provides insight in the complex relationships between the various aspects of social and economic life. Therefore, in an environment in which so many data sources on social issues are available, there is a great stimulus and challenge to *extend the scope* and to *improve the quality of social statistics* (Al and Bakker, 2000).

One great advantage of the social data sources nowadays is the availability of linkage-keys in the data files. There is no need anymore to present statistical information from isolated sources. On the contrary, the possibility to link several data files and to combine all kinds of information out of it, adds a new dimension to the processing of social statistics. For example, one can nowadays get information on employees (source: jobs register) who are married and are of non-European descent (source: PR). There is also a *value added* in the sense that it is possible to get *more reliable* and *complete* information when there are two or more sources with respect to the same subject.

4.1 Micro-linkage

Most of the present administrative registers are provided with a unique *linkage key*. It is the so-called *social security and fiscal number (SoFi-number)*, a personal identifier for every (registered) Dutch inhabitant and those abroad who receive an income from the Netherlands and have to pay tax over it to the Dutch fiscal authorities.

To prevent misuse of the SoFi-number, Statistics Netherlands recodes it for statistical processing into a so-called *Record Identification Number (RIN-person)*. Personal identifiers, such as date of birth and address, are replaced by age at the reference date and RIN-address. This is all done in accordance with regulations of the Dutch Data Protection Authority to *protect the privacy* of the citizens.

Since the SoFi-number is in use by social security administrations and tax authorities, one may expect it to be of excellent quality. A limited amount of SoFi-numbers may be registered with incorrect values in the data files, in which case linkage with other files is doomed to fail. However, in general, the percentage of matches is close to one hundred percent. Abuse of SoFi-numbers, for example by illegal workers, may occur in some cases, which results in a false match. Sometimes there are indications of a mismatch. An example of this is when the jobs register and the PR are linked and the worker turns out to be an infant. Another example is, when the FiBase shows an unusually high income for a worker, when it is in fact the sum of the incomes of all people using the same SoFi-number.

All social statistics data files can be linked to the PR. In practice this means that all these data files are indirectly linked to each other via the PR. Therefore the PR can be considered the backbone in the set of social data sources. When linking the PR and the jobs register, or the PR and a register of social benefits, it is a linkage between different statistical units (persons, jobs, benefits). One should realize in that case multiple *(1:n) linkage relationships* can exist because someone can have more than one job or can benefit from several social benefits.

In household sample surveys, like the LFS, records do not have a SoFi-number. For those surveys an alternative linkage key is used, which is often built up by a combination of the following personal identifiers: 1*) sex*; 2) *date of birth*; 3) *address* (in fact the combination of a postal code, mostly related to the street and house number. The postal code consists of four figures, followed by two letters). This sort of linkage key will usually be successful in distinguishing people. However, it is not a 100 percent unique combination of identifiers. Linking may result in a mismatch in the case of twins of the same sex. False matches may also occur when part of the date of birth or the

postal code and house number is unknown or wrong. Another drawback is that the linkage key is in fact not person but address related, which may cause linkage problems if someone has recently moved.

When linking the PR and the LFS with the alternative key, and tolerating a variation between sources in a maximum of one of the variables sex, year of birth, month of birth or day of birth, the result is that 96 to 97 percent of the LFS-records will be linked.

In its linkage strategy, Statistics Netherlands tries to *maximize the number of matches* and to *minimize the number of mismatches*. So, in order to achieve a higher linkage rate, more efforts are made to link the remaining unlinked records by means of different variants of the linkage-key. For example, leaving out the house number and tolerating variations in the numeric characters of the postal code. To keep the probability of a mismatch as small as possible, some 'safety' devices are built in the linkage process. This last linking attempt accomplishes an extra half percent matches.

In the end about two to three percent of the LFS records could not be linked to the PR. All together this is a good result, but *selectivity in the micro-linkage process* is not to be ruled out. If the unlinked records belong to a selective subpopulation, then estimates based on the linked records may be biased, because they do not represent the total population. Analysis in the past has indicated that the young people, in the 15-24 age bracket, show a lower linkage rate in household sample surveys than other age groups. The reason for this is that they move more frequently, therefore they are often registered at the wrong address. The linking rate for persons living in the four large cities Amsterdam, Rotterdam, The Hague and Utrecht is lower than for persons living elsewhere. Ethnic minorities also have a lower linkage probability, among other things because their date of birth is often less well registered (Arts et al., 2000).

From March 2004 the PR is going to serve as a sampling frame for the LFS. The prospect will be a matching rate of almost 100 percent, and no more linkage selectivity problems will occur.

4.2 Micro-integration

Successfully linking the PR with all the other data sources mentioned, makes much more coherent information on various demographic and socio-economic aspects of each individual's life available. One has to keep in mind, however, that some sources are more reliable than others. Some sources have a better coverage than others, and there may even be conflicting information between sources. So, it is important to recognise the strong and weak points of all data sources used.

Since there are differences between sources, there is a need for a *micro-integration* process to check data and to adjust incorrect data. It is believed that integrated data will provide far more reliable results, because they are based on an optimal amount of information. Also the coverage of (sub) populations will be better, because when data are missing in one source another source can be used. Another advantage of integration is that there is no reason for confusion among users of statistical information, because there will be *one figure on each social phenomenon*, instead of several figures depending on what source has been used.

During the micro-integration of the data sources the following steps have to be taken (Van der Laan, 2000a):

- harmonisation of statistical units;
- harmonisation of reference periods;
- completion of populations (coverage);
- harmonisation of variables, in case of differences in definition;
- harmonisation of classifications;
- adjustment for measurement errors, when corresponding variables do still not have the same value after harmonisation for differences in definitions;
- imputations in the case of item non-response;
- derivation of (new) variables; creation of variables out of different data sources;
- checks for overall consistency.

All these steps are controlled by a set of *integration rules* and are *fully automated*.

One example of how micro-integration works is the case in which data from the jobs register are confronted with data from the register of benefits. Both jobs and benefits are registered at volume base, which means that information on their state is stored at any moment in the year instead of at one reference day. Analysts of the jobs register know that the commencing date and the termination date of a job are not registered very accurately. It is important to know whether or not there is a job at the reference date, in other words whether or not the person is an employee. With the help of the register of benefits it is sometimes possible to define the job period more accurately.

Suppose that someone becomes unemployed at the end of November and gets unemployment benefits from the beginning of December. The jobs register may indicate that this person has lost the job at the end of the year, perhaps due to administrative delay or because of payments after job termination. The registration of benefits is believed to be more accurate. When confronting these facts the 'integrator' could decide to change the date of termination of the job to the end of November, because it is unlikely that the person simultaneously had a job and benefits in December. Such decisions are made with the utmost care. As soon as there are convincing counter indications of other jobs register variables, indicating that the job was still there in December, the termination date will in general not be adjusted.

5 The Social Statistical Database (SSD)

The micro-linkage and micro-integration of all the available data sources result in the end in the *Social Statistical Database (SSD)*, a whole *set of integrated micro-data files* in their definitive stage. The SSD contains coherent and detailed demographic and socio-economic statistical information on persons, households, jobs and (social) benefits. A major part of the statistical information is available on volume base. An extensive discussion on the SSD is found in Arts and Hoogteijling (2002).

The PR serves as *backbone* in this set of files (16 million records on 31 December 2000). Furthermore, the SSD contains the *integrated jobs file of employees* (7.4 million records on 31 December 2000) and the *integrated jobs file of self-employed persons*

(790 thousand records on 31 December 2000). Its sources were the EIS-Employees, the SEE, the FiTap and the FiBase. The jobs file is partially register and partially sample based. For example, a variable as 'time usually worked' comes from the SEE-sample and is only available for 3 million out of the 7.4 million records.

The main steps during the micro-integration process of the jobs file were:

- eliminating records (e.g. records applying to benefits instead of jobs; zero wage);
- merging records (e.g. different records probably belonging to the same job);
- adjusting date of commencement or termination (e.g. as a result of confronting jobs and benefits, see example above);
- adjusting wage;
- completion of missing data;
- addition of variables (from one source to another source where they are missing).

The SSD also has a set of integrated files of benefits, final EIS-UI-file (440 thousand records in 2000), final EIS-DI-file (1 million records in 2000), final SABA-file (580 thousand records in 2000) and a final file of 'other benefits' (5.6 million records in 2000) of which most records originate from the FiBase, including pensions and life insurances benefits. Apart from completion, the FiBase also plays a role in adjusting records in the integration process of benefits. The steps worth mentioning in integrating the files of benefits are:

- eliminating records (when all sources indicate that no sum of money has been paid out);
- adjusting date of commencement or termination (in case of administrative inaccuracy, for example after confronting with jobs);
- completion of missing data.

In trying to imagine what the SSD looks like, one should not think of a large-scale file with millions of records and thousands of variables. It would be very inefficient to store the integrated data in such a way. Also the issue of data protection prevents Statistics Netherlands from keeping so much information together. Instead, all the integrated files in their final stage are kept separately. There is just one combining element which is the linkage key RIN-person, present in every integrated file. So, whenever users demand a selection of variables out of the SSD-set, only the files with the variables demanded will be supplied. These can easily be extracted from the set and linked by means of the linkage key.

6 A SSD-based census

Much of the required census information deals with demographic aspects, and is supplied by the PR, the backbone of the SSD. Data on the employed population is extracted from the integrated jobs file (employees and self-employed persons).

Information on the retired population is partly obtained from the integrated file of 'other benefits' (pensions and life insurance benefits). For the remaining part of the Census Table Programme, information that cannot be found in registers, the LFS-file (the joint 2000 and 2001 file) is the main supplier of data.

Before census tabulation Statistics Netherlands needs to do some data processing, such as harmonisation and derivation, to make the SSD-data meet the Census Programme guidelines. Census tabulating is simply a matter of straightforward counting in the case of register variables. However, as soon as sample survey variables are involved, the method of 'repeated weighting' has to be applied in order to get consistent estimates between sample estimates and register totals.

Figure 1 gives an overview of the datasets that are used for the Census estimates. The horizontal axis presents the variables and the vertical axis presents the records.

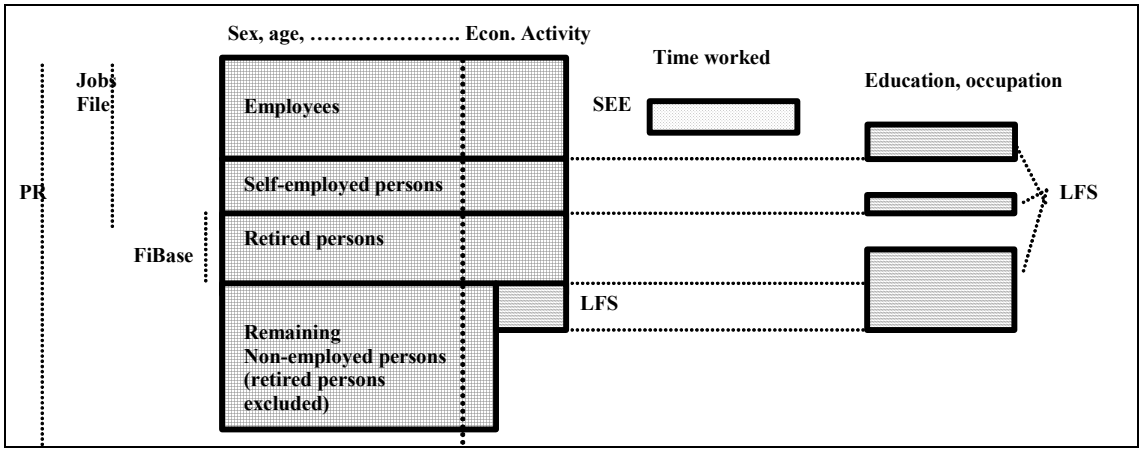

Figure 1: Micro-dataset from the SSD used for the Census 2001 estimation

6.1 SSD-harmonisation with Census 2001 guidelines

To make the SSD-data meet the Census 2001 guidelines, statistical units and reference periods have to be harmonised and census variables have to be derived.

- *Harmonisation of statistical units*: the population unit in the PR is the individual person, which is also the level demanded in most tables of the Census 2001 Programme. However, some of the census tables refer to the private household level. Therefore a household version of the PR had to be derived. Since, the household status is known of almost every person in the PR, as well as the household the person belongs to, this can easily be achieved by aggregation of PR-records from person level to household level.

  The relevant statistical unit in labour issues is the employee and not the job. If an employee has more than one job, it has been decided to refer to the characteristics of the main job. So, if the branch of economic activity and the working hours of an employee are to be tabulated, those of the main job are taken. The main job is defined as the job with the highest gross wage.

- *Harmonisation of reference periods*: to select the data from the SSD in accordance with the reference date of the Dutch Census, which is 1 January 2001, one has to harmonise the reference periods.

  This means that demographic data in the PR are selected according to the situation on 1 January 2001.

  Reference periods in the case of the jobs file are harmonised by making a selection of employees with a job (of at least 1 hour a week) in the period of 22-31 December 2000. The reasons for a slight deviation of the Census reference date are the following. First, information on jobs at 1 January 2001 was not yet available. Second, every year the job patterns show a dip in the last week of December. It is likely that many flexible workers jobs terminate before the end of the year. Therefore, 31 December is less appropriate as a choice for a representative reference date.

  To get the number of self-employed persons at Census reference date we take the complete file of self-employed persons, since no reference period is known of their jobs.

  The retired population in the age class of 55-64 are selected by tracing persons in the FiBase who have an income exclusively from pensions or life insurance benefits at the end of the year 2000. The age variable is measured at 1 January 2001.

  The relevant LFS-variables for the Census are educational attainment level, occupation, and a variable 'current activity', which indicates whether a person is unemployed, attending full-time education or is engaged in family duties. As stated before, the LFS is a continuous survey, which is conducted throughout the year. Therefore, the LFS-information on the survey date will not always be valid at the Census reference date of 1 January 2001. Suppose the LFS finds that someone is unemployed at the survey date. If the integrated jobs file indicates that this person has a job at the Census reference date, he or she will be qualified as employed for the Census. This means that the information from the integrated jobs file overrules the LFS-information. It prevents the incorrect classification of the person as unemployed at the reference date.

- *Derivation of census variables*: the Census Programme guidelines (Eurostat, 1999) are not very clear on how to deal with persons with a mix of (economic) activities. Therefore Statistics Netherlands defined some priority rules to make an unambiguous choice in (economic) activity. The census variable '(economic) activity of a person' is defined in the following way:
  1) People from 0 to 3 years old are by definition *other economically inactive*;
  2) People from 4 to 15 years old are by definition *attending full-time education*, even if they have a job;
  3) People of 75 years and older are by definition *retired*; this also applies to the few persons aged over 75 who are still working;
  4) People from 65 to 74 and without job are by definition *retired*;
  5) People from 55 to 64 and with an income exclusively from pension or insurance benefits for retirement purposes are by definition *retired*;
  6) People in the ages 16-74 who have an employee job (of at least 1 hour a week) are 'economically active' and have the employment status *employee*;

7) People in the ages 16-74 who have a job as a self-employed person and no employee job are 'economically active' and have the employment status *self-employed person*.

What remains is the population in the ages 16-74 that is neither employed nor retired. Some of them are *economically active although unemployed*. Others are *economically inactive*, but they are active in another sense, either *attending full-time education* (restricted to persons of at most 30 years old), being *engaged in family duties* or being *other economically inactive*. No register information is available to distinguish between all these activities, only LFS information.

6.2 Census counts, SSD-register variables

When it comes to tabulating SSD-*register variables*, it is just a matter of straightforward counting of the register data. Register counts from the SSD will always be numerically consistent in all census tables. This is guaranteed because the SSD-database consists of micro-integrated files in which conflicting information has been harmonised.

Table 1 shows some key figures from the Census 2001 on the population aged 0-14, 15-74 and 75+. The numbers were counted from the register part of the SSD.

Table 1 : Key figures for the age groups 0-14; 15-74 and 75+ based on SSD-register part, Census 2001.

	Total population	**Age-group**		
		0-14	15-74	75+
Total population	15.985.538	2.977.283	12.036.171	972.084
Sex				
Males	7.909.052	1.522.811	6.047.425	338.816
Females	8.076.486	1.454.472	5.988.746	633.268
Country of birth				
The Netherlands	14.370.439	2.843.537	10.616.115	910.787
Other European Union countries	307.723	26.015	257.541	24.167
Non-European Union countries	1.307.376	107.731	1.162.515	37.130
Employed population				
Total	7.394.777	-	7.394.777	-
Employee	6.786.511	-	6.786.511	-
Self-employed person	608.266	-	608.266	-
Non-employed population				
Total	8.590.761	2.977.283	4.641.394	972.084
Retired	2.328.024	-	1.355.940	972.084
Other non-employed	6.262.737	2.977.283	3.285.454	-

Source: Schulte Nordholt (2003) and Population Census 2001.

6.3 Repeated weighting, SSD-survey variables

As soon as *survey variables* in a sample are involved, the census figures will have to be estimated. Estimating (sub) totals from the SEE or LFS samples consistent with the register totals is a rather complex issue.

Sample survey data are normally provided with sample weights, which are calibrated in such a way that table estimates become representative of the population. With these weights, all the variables are inflated in the same way. Estimates from a survey are always numerically consistent as long as they are based on the same micro data, and use the same weights. However, they are generally not numerically consistent with all register counts, except for the few register variables used as auxiliary information in the calibration model. One should realize that it is next to impossible to take into account all the register information, because that would give too many restrictions and lead to estimation problems. So it is impossible to realize full numerical consistency between sample estimates and register counts with the traditional way of weighting.

When users have to produce table sets for the census, they will demand *overall numerical consistency* between all tables in this table set. The need for overall numerical consistency stimulated methodologists at Statistics Netherlands to develop a new estimation method that ensures numerically consistent table sets, even if the data are obtained from different data sources. The method is called *repeated weighting (RW)*; it is based on the repeated application of the regression method to eliminate numerical inconsistencies between table estimates from different sources.

Roughly speaking, the RW method works as follows. Each table is estimated using as many records as possible: depending on the variables of interest, the table may be counted from register data, or estimated from survey data from one or more surveys. Then for each table, one determines which margins the present table has in common with the tables in the set that are already estimated. The next step is to estimate the table while calibrating on these common margins. Apart from being consistent, the estimates will also be more accurate, particularly if the margins can be estimated from larger data sets or counted from register data. As such these margins serve as auxiliary information.

Whereas with traditional weighting one fixed set of weights is calculated per sample survey, with the RW method one derives a new set of weights (based on the survey weights) per table in order to get consistency with tables that are already estimated.

Those who want to know more about the principles of the RW-method should read the position paper of Houbiers et al. (2003).

An important prerequisite in the preparation to repeated weighting estimation is that internal consistency of all the relevant integrated micro-data must be guaranteed. The RW method cannot solve inconsistencies in the data that are caused by differences in definition or by measurement errors and that were not repaired in the micro-integration process. If such inconsistencies show up they have to be repaired first. The purpose of repeated weighting is to correct for sample errors, and it is certainly not a micro-integration tool.

Empty cells, with no survey observations at all, may cause estimation problems. Particularly so if there are persons in the population who belong to that cell according to the registers, while none of them is surveyed. Survey zeros occur often when little data is available and when tables have to be estimated at a detailed level.

Table 2 shows the results of repeated weighting on a selected set of sample variables in the LFS and compares them with initial LFS-weight results.

Table 2 : A selection of estimated sample variables for persons of 15-74 years old, 2001

	LFS 2000-2001 Initial LFS-weights	Census 2001 RW
	In % of population of 15-74 years old	
Total population 15-74 years old	100,0	100,0
Educational attainment		
ISCED 0+1 (primary level or less)	15,8	16,5
ISCED 2+3+4 (secondary level)	64,9	64,7
ISCED 5+6 (tertiary level)	19,3	18,8
Occupation		
ISCO-COM 1 (legislators, senior officials, managers)	8,1	7,7
ISCO-COM 2 (professionals)	10,7	10,0
ISCO-COM 3 (technicians and associate professionals)	11,1	10,4
ISCO-COM 4 (clerks)	7,5	7,0
ISCO-COM 5 (service workers, shop and market sale workers)	7,1	6,7
ISCO-COM 6 (skilled agricultural and fishery workers)	0,9	0,9
ISCO-COM 7 (craft and relative workers)	6,2	5,6
ISCO-COM 8 (plant and machine operators and assemblers)	4,0	3,7
ISCO-COM 9 (elementary occupations)	4,6	4,3
ISCO-COM 0 (armed forces)	0,3	0,3
Occupation unknown	4,8	4,6
Non-employed	34,8	38,6

Source: LFS 2000, LFS 2001 and Population Census 2001

When analyzing the results on educational attainment in table 2, one can see that by repeated weighting the LFS-outcome of the lower education level (ISCED 0+1) is adjusted upward, whereas the higher education level (ISCED 5+6) is adjusted downward. Apparently the LFS-weighting does not contain enough auxiliary information to correct for selective non-response in the education levels.

Why is so much effort invested in such an ambitious project as estimation by repeated weighting, instead of simply applying *mass imputation* on all the records with missing information? An important advantage of mass imputation is that once the records are imputed for missing data, users estimating with this file will always be able to reproduce results. Statistics Netherlands has gained a lot of experience with mass imputation. The main reason not to go on with the practice of mass imputation is that the imputation models can never be sufficiently rich to account for all significant data patterns between sample and register data. There is also a danger that it may lead to oddities in the estimates (Kooiman, 1998). A simple example of this is the case in which a file is built up out of register information, including the variable age, and of sample information from the LFS, including education level. Suppose, all non-sample records are mass-imputed for missing data on education levels. A researcher may conclude on the basis of this mass-imputed file that there are highly educated babies. Such things can happen if the relationships between age and education level were not taken account.

7 Concluding remarks

Statistics Netherlands has innovated its methods of data collecting and data processing for the compilation of the Census Table Programme 2001. The most important elements in the new approach are the use of a combination of administrative registers and household sample surveys as data sources, and the application of repeated weighting, a new methodology to estimate numerically consistent tables from this data source. The result is called a virtual census, because the results for some characteristics of the population are based on estimates instead of enumeration.

The new way of producing census tables proved to be a succesful and much cheaper alternative for the costly census projects of the past. No special effort had to be made to collect data, as the data sources used for the 2001 Census were already part of the regular statistical programme of Statistics Netherlands. Most data came from registers, and only some supplementary information was needed from sample surveys. This means that the implementation of the 2001 Census Programme hardly caused any response burden. Moreover the data processing time for the 2001 Census was just a fraction of what it would have been in a traditional census.

One disadvantage of the new approach is that Statistics Netherlands was not capable of fully publishing tables that have cells with little or even no sample survey data. Some estimates were considered too unreliable to publish. And, in the zero-cell case, it was impossible to do so. This applies in particular to the census tables that demand detailed information for small subpopulations, which is certainly not a hypothetical case. The small sample problem shows up in tables in the Census Programme 2001 that are specified at a detailed regional level, such as municipalities. A traditional census would not have such a problem.

So, the problem of low cell size will have to be solved, for example, by oversampling small subpopulations. Increasing the sample size is not always sufficient, because in the actual case even the joint LFS 2000 and 2001 could not prevent low cell sizes in all tables. Another solution for the small sample problem would be to develop administrative registers with data that are now obtained by sample surveys. For example, Statistics Netherlands is planning to build a register in which the level of educational attainment will be stored for all citizens. This is a long-term project though. Another possibility would be the use of so-called small area estimators.

When one compares the present way of compiling census tables with those used in the Censuses of 1981 and 1991, the Census Programme 2001 may have required some more production time but the estimates in cross-tabulations of register and survey information are more accurate. First, because the statistical information has gained much more coherence because of combining the data sources. Second, because more auxiliary information could be used in the estimation methods than in the past since more registers are available. Third, because much effort has been made to achieve overall numerical consistency.

The experiences with the new approach proved so convincing, that it is absolutely worth while to build on these experiences for the Census Programme of 2011. Besides, the policy of Statistics Netherlands is to extend its data collection with more administrative sources.

In Summer 2004 a book on the Dutch Census 2001 is to appear in English (Statistics Netherlands, 2004).

Appendix 1. Overview of data sources Census 2001

Table 3: Overview of the data sources of the 2001 Census Programme

	Source	Statistical unit	Integral or sample/ number of records in reference period	Reference period	Census variables
1.	Population Register (PR)	Person	Integral/ 16.0 million records	1 January 2001	• Sex • Age • Country of birth • Country of citizenship • Country of residence a year prior to census • Region of residence • Marital status • Family status • Family situation • Family nucleus type • Private household composition • Household status • Household size • Number of children • Other persons in household (outside the family nucleus)
2.	Employee Insurance Schemes Registration System – Employees (EIS-Employees)	Job	Integral/ 6.5 million records	22-31 December 2000	• Employee • Branch of economic activity (NACE) • Gross wage before deduction of social insurance premiums (auxiliary variable)
3.	Survey on Employment and Earnings (SEE)	Job	Selective sample/ 3.0 million records	22-31 December 2000	• Employee • Time usually worked • Place of work
4.	Register of final income tax assess-ments on profits of self-employed persons (FiTap)	Self-employed person	Integral/ 790 thousand records (missing about 5 percent records)	Volume 2000	• Self-employed person • Branch of economic activity (NACE)
5.	Employee Insurance Schemes Registration System – Unemployment insurance (EIS-UI)	Benefit	Integral/ 440 thousand records	Volume 2000	No census variables
6.	Employee Insurance Schemes Registration System – Disablement insurance (EIS-DI)	Benefit	Integral/ 1.0 million records	Volume 2000	No census variables
7.	Social Assistance Benefits Administration (SABA)	Benefit	Integral/ 580 thousand records	Volume 2000	No census variables
8.	FiBase-register (register of advance tax payments)	Income transaction	Integral/ 7.2 million records (jobs employees); 2.7 million records (pensions/ life insurance)	31 December 2000	• Employee: yes/no. • Pensions / life insurance benefits (retired persons)
9.	Labour Force Survey (LFS)	Person	Sample/ 2000: 120 thousand records; 2001: 110 thousand records	2000 and 2001	• Educational attainment level • Occupation • Unemployment • Attendance at educational institutions • Engaged in family duties

This overview is partially extracted from Van der Laan (2000b), page 23.

References

P.G. Al and B.F.M. Bakker. Re-engineering Social Statistics by Micro-integration of Different Sources, An Introduction. *Netherlands Official Statistics*, Vol. 15 (Summer 2000): Special Issue, *Integrating Administrative Registers and Household Surveys*, ed. P.G. Al and B.F.M. Bakker: pages 4 6, 2000.

C.H. Arts, B.F.M. Bakker and F.J. van Lith. 'Linking Administrative Registers and Household Surveys'. *Netherlands Official Statistics*, Vol. 15 (Summer 2000): Special Issue, *Integrating Administrative Registers and Household Surveys*, ed. P.G. Al and B.F.M. Bakker: pages 16 22, 2000.

C.H. Arts and E.M.J. Hoogteijling. The Social Statistical Database of 1998 and 1999. *Monthly Bulletin of Socio-economic Statistics*. Vol. 2002/12 (December 2002): pages 13-21, 2002 [in Dutch].

P. Corbey. Exit the population census. *Netherlands Official Statistics*, Vol. 9 (Summer 1994): pages 41-44, 1994.

Eurostat (Statistical Office of the European Communities). Guidelines and Table Programme for the Community Programme of Population and Housing Censuses in 2001. Vol. 2, Table Programme. Eurostat Working Papers, Population and social conditions No. 3/1999/E/No10. Luxembourg: Eurostat, 1999.

Eurostat (Statistical Office of the European Communities). *Documentation of the 2000 Round of Population and Housing Censuses in the EU, EFTA and Candidate Countries*, ed. B. Kotzamanis. Prepared on behalf of Eurostat by Laboratory of Social and Demographic Analysis (LDSA), University of Thessaly (Volos), Greece, 2003.

C. Harmsen and A. Israëls, 2003. Register-based Household Statistics. Paper prepared for *the European Population Conference 2003: European Populations: Challenges and Opportunities*, Warsaw, Poland, 26-30 August 2003.

E.M.J. Hoogteijling. Illegal people in the Netherlands. *Monthly Bulletin of Population Statistics*. Vol. 2002/03 (March 2002), page 21, 2002 [in Dutch].

Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen and V. Snijders. *Estimating consistent table sets: position paper on repeated weighting*, Discussion paper 03005, Voorburg/Heerlen: Statistics Netherlands, 2003. See: http://www.cbs.nl (select English/ publications/articles/discussion papers).

P. Kooiman. *Mass imputation*: Why not!?, Research paper, BPA-no. 8792-98-RSM, Statistics Netherlands, Voorburg, 1998 [in Dutch].

P. van der Laan. Integrating Administrative Registers and Household Surveys. *Netherlands Official Statistics*, Vol. 15 (Summer 2000): Special Issue, Integrating *Administrative Registers and Household Surveys*, ed. P.G. Al and B.F.M. Bakker, pages 7 15, 2000a.

P. van der Laan. The 2001 Census in the Netherlands: Integration of Registers and Surveys. Paper prepared for the *INSÉÉ-EUROSTAT Seminar on Censuses After 2001*, Paris, France, 20 and 21 November 2000, 2000b.

F. Linder. The Dutch Virtual Census 2001, a new approach by combining Administrative Registers and Household Sample Surveys. Paper prepared for the *DIECOFIS Workshop on Data Integration and Record Matching*, Vienna, Austria, 13-14 November 2003.

C.J.M. Prins. Dutch population statistics based on population register data. *Monthly Bulletin of Population Statistics*. Vol. 2000/02 (February 2000), pages 9-15, 2000.

N.E. Schulte. The "Virtual Census" 2001, paper prepared for the *Statistics Netherlands-SISWO workshop on the Virtual Census and the SSD*, Amsterdam, 11 November 2003.

Statistics Netherlands. *Censuses in the twentieth century. Research methods on Households and Enterprises*. J.G.S.J. van Maarseveen (ed.), J.G.S.J. van, pages 5-142, 2002 [in Dutch].

Statistics Netherlands. *The Dutch Virtual Census of 2001, Analysis and Methodology*, M.B.G. Gircour, M.I. Hartgers and E. Schulte Nordholt (ed.), 2004 (forthcoming).

J.M. Vliegen and P. van der Laan. The "Census" in the Netherlands: integration of register and sample survey data. *Sonderhefte zum Allgemeinen Statistischen Archiv. Organ der Deutschen Statistischen Gesellschaft*, Heft 33: Volkszählung 2001. Von der traditionellen Volkszählung zum Registerzensus, ed. H. Grohmann, H. Sahner and R. Wiegert, pages 15-23, 1999 [in German].

J.M. Vliegen and H. van de Stadt. Is a Census still necessary? Experiences and alternatives. *Netherlands Official Statistics*, Vol. 3 (3), pages 27-34, 1988.

Author's address:

Frank Linder
Statistics Netherlands
Division for Social and Spatial Statistics
Unit Statistical Analysis
P.O. Box 4000
2270 JM Voorburg
The Netherlands

Tel. +31 703 375 / 711
Elec. Mail: flnr@cbs.nl
http://www.cbs.nl