2007

virtual censuses

zentsu birtualak

censos virtuales

ERIC SCHULTE NORDHOLT

47

nazioarteko estatistika mintegia euskadin

seminario internacional de estadística en euskadi

Lanketa / Elaboración:

Euskal Estatistika Erakundea
Instituto Vasco de Estadística (EUSTAT)

Argitalpena / Edición:

Euskal Estatistika Erakundea
Instituto Vasco de Estadística
Donostia – San Sebastián, 1 – 01010 Vitoria – Gasteiz

Euskal AEko Administrazioa
Administración de la C.A. de Euskadi

Ale-kopurua / Tirada:
500 **ale** / ejemplares

XI-2007

Inprimaketa eta Koadernaketa*:*
Impresión y Encuadernacion:
Estudios Gráficos ZURE S.A.
Ctra. Lutxana-Asua, 24 A
Erandio-Goikoa (BIZKAIA)

I.S.B.N.: 978-84-7749-445-4
Lege-gordailua / Depósito Legal: BI-3717-07

AURKEZPENA

Nazioarteko Estatistika Mintegia antolatzean, hainbat helburu bete nahi ditu EUSTAT-Euskal Estatistika Erakundeak:

- Unibertsitatearekiko eta, batez ere, Estatistika-Sailekiko lankidetza bultzatzea.
- Funtzionarioen, irakasleen, ikasleen eta estatistikaren alorrean interesatuta egon daitezkeen guztien lanbide-hobekuntza erraztea.
- Estatistika alorrean mundu mailan abangoardian dauden irakasle eta ikertzaile ospetsuak Euskadira ekartzea, horrek eragin ona izango baitu, zuzeneko harremanei eta esperientziak ezagutzeari dagokienez.

Jarduera osagarri gisa, eta interesatuta egon litezkeen ahalik eta pertsona eta erakunde gehienetara iristearren, ikastaro horietako txostenak argitaratzea erabaki dugu, beti ere txostengilearen jatorrizko hizkuntza errespetatuz; horrela, gai horri buruzko ezagutza gure herrian zabaltzen laguntzeko.

Vitoria-Gasteiz, 2007ko Azaroa

JOSU IRADI ARRIETA
EUSTATeko Zuzendari Nagusia

PRESENTATION

In promoting the International Statistical Seminars, EUSTAT-The Basque Statistics Institute wishes to achieve several aims:

- Encourage the collaboration with the universities, especially with their statistical departments.
- Facilitate the professional recycling of civil servants, university teachers, students and whoever else may be interested in the statistical field.
- Bring to the Basque Country illustrious professors and investigators in the vanguard of statistical subjects, on a worldwide level, with the subsequent positive effect of encouraging direct relationships and sharing knowledge of experiences.

As a complementary activity and in order to reach as many interested people and institutions as possible, it has been decided to publish the papers of these courses, always respecting the original language of the author, to contribute in this way towards the growth of knowledge concerning this subject in our country.

Vitoria-Gasteiz, November 2007

JOSU IRADI ARRIETA
General Director of EUSTAT

PRESENTACIÓN

Al promover los Seminarios Internacionales de Estadística, el EUSTAT-Instituto Vasco de Estadística pretende cubrir varios objetivos:

- Fomentar la colaboración con la Universidad y en especial con los Departamentos de Estadística.
- Facilitar el reciclaje profesional de funcionarios, profesores, alumnos y cuantos puedan estar interesados en el campo estadístico.
- Traer a Euskadi a ilustres profesores e investigadores de vanguardia en materia estadística, a nivel mundial, con el consiguiente efecto positivo en cuanto a la relación directa y conocimiento de experiencias.

Como actuación complementaria y para llegar al mayor número posible de personas e Instituciones interesadas, se ha decidido publicar las ponencias de estos cursos, respetando en todo caso la lengua original del ponente, para contribuir así a acrecentar el conocimiento sobre esta materia en nuestro País.

Vitoria-Gasteiz, Noviembre 2007

JOSU IRADI ARRIETA
Director General de EUSTAT

BIOGRAFI OHARRAK

ERIC SCHULTE NORDHOLT Statistics Netherlands erakundeko ikertzaile titularra eta proiektuburua da. Matematikan eta Ekonometrian graduatu eta gero, Statistics Netherland-en sartu zen 1992an. Hasieran estatistika-metodologiaren sailean aritu zen, eta 1996az geroztik gizarte-estatistiken arloan. 1995ean Eurostatekin elkarlanean aritu zen Luxemburgen, eta 2006an Statistics New Zealand-ekin. Eric Schulte Nordholt-ek eskarmentu handia du estatistika-ikastaroetako irakasletzat, ikastaro orokorretan nahiz espezializatuetan. Holandako zentsu birtualaren arduraduna da, non taula guztiak lehendik dauden informazio-iturrietan oinarrituta (erregistroak eta inkestak) kalkulatzen dituzten. Hori eginez, Holandak ohiko zentsuen ordez bestelako irtenbide merkea eta fidagarria aurkitu du. Eric Schulte Nordholt hainbat erakundetako kide da, besteak beste, ISI, IAOS, IASS eta Estatistikak eta Eragiketak ikertzeko Holandako Erakundekoa.

BIOGRAPHICAL SKETCH

ERIC SCHULTE NORDHOLT is a senior researcher and project leader at Statistics Netherlands. After graduating in Mathematics and Econometrics, he joined Statistics Netherlands in 1992. He first worked in the department of Statistical Methods and since 1996 he works in Social Statistics. In 1995 he had a secondment at Eurostat in Luxembourg and in 2006 at Statistics New Zealand. Eric Schulte Nordholt has a lot of experience in teaching both general and specialist statistical courses. He is responsible for the Dutch Virtual Census where all tables are estimated based on already existing data sources (registers and surveys). This way, the Netherlands found a cheap and reliable alternative for the traditional Censuses. Eric Schulte Nordholt is a member of the ISI, IAOS, IASS and Netherlands Society For Statistics and Operations Research.

NOTAS BIOGRÁFICAS

ERIC SCHULTE NORDHOLT es investigador titular y jefe de proyectos en Statistics Netherlands. Tras graduarse en Matemáticas y Econometría, ingresó en Statistics Netherlands en 1992. Inicialmente trabajó en el departamento de Metodología Estadística y desde 1996 en Estadísticas Sociales. En 1995 colaboró con Eurostat en Luxemburgo y en 2006 con Statistics New Zealand. Eric Schulte Nordholt tiene una amplia experiencia como profesor de cursos de estadística, tanto general como especializada. Es responsable del Censo Virtual Holandés donde todas las tablas son estimadas basándose en fuentes de información ya existentes (registros y encuestas). De esta manera Holanda ha encontrado una alternativa barata y fiable a los Censos tradicionales. Eric Schulte Nordholt es miembro del ISI, IAOS, IASS y de la Sociedad Holandesa para la Investigación de Estadísticas y Operaciones.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

The Dutch Virtual Census:

Combining data from registers and sample surveys

SUMMARY

Data from many different sources were combined to produce the census 2001 tables for the Netherlands. Since the last census based on a complete enumeration was held in 1971, the willingness of the population to participate fell sharply. Statistics Netherlands found an alternative in the virtual census, using registers and surveys already available. The virtual census is cheaper, comparable to earlier Dutch censuses and more socially acceptable. The Netherlands takes up a unique position in the European Census Round. The table results are comparable with the earlier Dutch censuses and with those of the other countries in the 2001 Census Round.

In producing tables for the 1981 and 1991 census rounds, the focus in the Netherlands was on combining population register data with results from the Labour Force and Housing Surveys. Less care was given to overall consistency of the table estimates. For the 2001 census Eurostat and other international organisations required more detailed information than for earlier Census Rounds. Moreover, in the last decade Statistics Netherlands has acquired more and more experience in dealing with data of various administrative registers for statistical use. This enabled the development of a Social Statistical Database (SSD), which contains coherent and detailed demographic and socio-economic statistical information on persons and households. The Population Register forms the backbone of the SSD. The SSD is constructed by micro linking several administrative registers and sample surveys. A micro integration process ensures coherence, consistency and completeness of the SSD data. Sample surveys are still needed for information that is not available from registers. Examples of variables that are not available in the Dutch registers are level of education and occupation. However, these two variables are available in the Dutch Labour Force Survey.

Overall numerical consistency among all tables in the census tables set is required. This need stimulated methodologists at Statistics Netherlands to develop a new estimation method that ensures numerically consistent table sets if the data are obtained from different data sources. The method is called repeated weighting, and is based on the repeated application of the regression method to eliminate numerical inconsistencies among table estimates from different sources. The newly developed method of repeated weighting guarantees that combining survey and register information leads to consistent estimates in the tables of the Census Programme. The Dutch census tables 2001 have been estimated by making use of all available information and therefore the best possible quality of the estimates has been obtained.

Remarks:
The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

In this monograph the availability of Dutch census microdata from 1960, 1971 and 2001 is described in more detail. Although some harmonisations between countries and census years have been reached, some recommendations are done to improve the future comparability.

In the last section of this monograph some background information is given about Statistical Disclosure Control (SDC). SDC is one of the methodological aspects of the statistical production process. The aim is of course to release as much census information as possible. However, the privacy of individual respondents should be respected. Therefore, Statistical Disclosure Control techniques have been developed to protect sensitive information that can be attributed to individual respondents. SDC is thus relevant to be able to decide properly what kind of census tables and microdata can be released.

Keywords: census; consistent table estimates; micro datasets; micro integration; micro linking; repeated weighting; Social Statistical Database (SSD); Statistical Disclosure Control (SDC)

1. INTRODUCTION

In 2003 data were combined to produce the Dutch 2001 census tables. In the Netherlands this was not done by interviewing inhabitants in a complete enumeration, but by using data that Statistics Netherlands already had available. This way, the Dutch tax payer got a much lower census bill. The costs for a traditional census would be about three hundred million euros, while the costs made now are 'only' about three million. The estimate includes the costs for all preparatory work such as developing a new methodology and accompanying software. The costs of the registers are not included, but the analyses of the results are. Registers are not kept up-to-date for censuses but for other purposes. Saving money on census costs is only possible in countries that have sufficient register information. As an example we can compare the costs of the Dutch virtual census with the costs of the traditional census that was held in Canada. In Canada the census costs amounted to approximately 450 million euros. Canada has about 31.6 million inhabitants, twice as many as the Netherlands. Statistics Canada justifies the huge census costs by pointing out the enormous implications of the census results for the distribution of money among regions. Moreover, a virtual census would be impossible in Canada because of the lack of sufficient register data.

The 2001 census relates to forty extensive tables. Twenty-eight are about the Netherlands as a whole, nine are at the COROP level (NUTS 3) and three at municipal level (NUTS 5). The forty tables fall into a number of groups. Eight tables concern housing, two tables concern commuting and the other thirty tables are demographic tables, relating to occupation, level of education and economic activity. Additionally, demographic, housing and labour figures are compiled at sub-city district level for ten large cities that participate in Urban Audit II (Statistics Netherlands, 2003). These ten large cities are Amsterdam, Rotterdam, The Hague, Utrecht, Eindhoven, Tilburg, Groningen, Enschede, Arnhem and Heerlen.

Except the financial aspect, other important differences exist between a traditional census and the virtual census conducted in the Netherlands. In spite of the mandatory character of a traditional census, a certain part of the population will not participate (unit non-response) and the part that does participate will not answer some questions (item non-response). Correcting non-response by weighting and imputation techniques is well worth trying. A well-known problem with traditional censuses is that participation is limited and selective. Traditional correction methods fall short of the need to be able to publish reliable results. The last traditional census in the Netherlands (in 1971) met with much privacy objections against the collection of integral information about the population living in the Netherlands. This increased the non-response problem and the expectation was that non-response would be even higher if another traditional census were held in the Netherlands (Corbey, 1994). There are almost no objections to a virtual census and the non-response problem only plays a role in the

surveys of which the data are used. If non-response can be corrected in a survey, it will certainly be possible to correct for the selectivity of that survey in the census where it is used.

The virtual census in the Netherlands was off to a later start than in other countries where a traditional census was conducted. It did not make sense to really start the 2001 Census Project until all sources were available; some registers were available relatively late. Nevertheless, the Netherlands was quicker with the compilation of the forty census tables than most of the other countries that participated in the 2001 Census Round. In fact, the Netherlands was one of the first to send the complete set of forty tables to Eurostat, which co-ordinated the contributions of all European Union (EU) member states, accession countries and European Free Trade Association (EFTA) member states. The Netherlands had the advantage that the incoming census forms did not need to be checked and corrected. However, one must realise that for some variables only sample information is available, which implies that it was impossible to meet the level of detail required in some Dutch tables.

An interesting option for the future is to use small area estimation techniques to estimate the cell values that could not be estimated adequately. A theoretical framework for small area estimation can be found in Rao (2003). The ONS studied the application of this technique in the context of its Neighbourhood Statistics Programme. This is a major initiative to bring together and make widely available statistics on a small area level. In each case of implementation of indirect small area estimates particular attention was paid to model specification. Some experimental synthetic estimates were published in the United Kingdom and others are undergoing a process of evaluation. Possibly, the techniques of repeated weighting and small area estimation can be combined in the 2011 Census Round.

Currently, the advantages of the virtual census in cost and non-response problems amply make up for the loss of some detail compared to a traditional census. Moreover, not all information required will always be available for the users in traditional censuses. This is because traditional correction methods such as weighting and imputation sometimes do not correct for limited and selective participation. This means no reliable results can be published for some of the cells in the set of tables. One may wonder why simply applying mass imputation (filling in valid values for all missing scores) is not taken into account to overcome these problems. An important advantage of mass imputation is that once the records are imputed, any user will be able to reproduce results when using the same imputed file. However, mass imputation is not a viable strategy for raising survey outcomes to population totals. There are not enough degrees of freedom to sustain a sufficiently rich imputation model accounting for all significant data patterns between sample and register variables. Only if the interest is in totals of subsets of the population defined by the explanatory variables in the model, the imputation approach leads to approximately design-unbiased and hence reliable estimates (at least if the variances are reasonably small) (Kroese and Renssen, 2000).

The Nordic countries (Denmark, Finland, Iceland, Norway and Sweden) have more variables available in registers than the Netherlands. So the problem of insufficient detail in the outcome does not play a major role there. Moreover, some Nordic countries conduct a

(limited) enumeration for variables missing in the registers. Most of the other countries are in a similar position as the Netherlands where some variables relevant for the census can be found in registers, while other variables are available on a sample basis only. That's why much interest exists in the Dutch approach to combine registers and surveys and to use modern statistical techniques and accompanying software to compile the tables. In September 2003 three colleagues of the ONS visited Statistics Netherlands to learn more about the Dutch approach. It is of course crucial that statistical bureaus may make use of registers that are relevant for the census. For Statistics Netherlands this was laid down in the new statistical law of 2003. Nevertheless, in the years to come Statistics Netherlands will have to establish good contacts with register holders. Timely deliveries with relevant variables for Statistics Netherlands are crucial for statistical production.

The reason why Statistics Netherlands has compiled the set of tables is a gentlemen's agreement. In 1991 the Census Act was rescinded, officially cancelling Statistics Netherlands obligation to hold a census once every ten years (Corbey, 1994). There is no European obligation to supply census data, but it is almost inconceivable that the Netherlands would not compile census data for the international organisations just like all other European countries do. Eurostat has a co-ordinating role in collecting harmonised data on the EU and a duty to make international comparisons of the outcome.

It will be several years before all countries participating in the 2001 Census Round sent their final set of tables to Eurostat. Therefore, Statistics Netherlands took the initiative to compare the 2001 results of a limited number of European countries. The results of the Dutch 2001 census were also compared to earlier Dutch censuses. Such work has been carried out in the past as well. The data compiled on 1981 and 1991 were much less detailed than the set of tables of the 2001 census. The 1991 Dutch census was largely based on a register count of the population in combination with the Labour Force Survey 1991 and the Housing Demand Survey 1989/1990. Contrary to 1981 and 1991, Statistics Netherlands has published census information over 2001 on the municipal level.

2. METHOD OF COMPILING, MICRO LINKAGE, MICRO INTEGRATION AND THE SOCIAL STATISTICAL DATABASE (SSD) AND REPEATED WEIGHTING

2.1 Method of compiling

The current virtual census relates to 2001. The backbone of the census is the central Population Register (PR), which is the combination of all municipal population registers. The Population Register (PR) contains demographic information on every inhabitant of the Netherlands (Prins, 2000).

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 when two persons are registered at separate addresses, but actually live together. They have a financial incentive to be registered at different addresses if one person is employed and the other is on welfare. 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. This 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 to be 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).

A number of integrated surveys and registers were linked to the PR. For this linking process only exact matches are used based on the unique so-called Social security and Fiscal (SoFi) number. The integrated system is called the Social Statistical Database (SSD) system. It is developed originally to conduct virtual censuses, but now it is also used for many social statistics.

PR data of 1 January 2001 were used as the basis for the set of tables. The set of tables focuses on frequency counts and not on quantitative information. The SSD datasets on 2001 (the Social Statistical Database include integrated microdata on employees and self-employed) were not available on time for the 2001 census. Therefore, we used datasets of

2000 that were available in the beginning of 2003 to deduce the individual data of the end of 2000 as an approximation for the situation on 1 January 2001.

Different variables, such as occupation and level of education, were obtained from the Labour Force Survey (LFS). The variable job size was obtained from the large Survey on Employment and Earnings (SEE). To obtain sufficient records, information on persons from the LFS 2000 and the LFS 2001 was combined. For the housing tables, we used PR data of 1 January 2001, the Housing Register 2001 and the Survey on Housing Conditions (SHC) 2000. For the tables on commuting we used the PR data of 2000 and 2001, the SEE 2000 and the SSD datasets of 2000.

Some variables of the PR and SSD datasets are available on an integral basis. Examples are age, sex, marital and employment status. Survey variables are only available for a part of the population. Examples are the highest level of education attained (LFS) and whether someone rents or owns the property they live in (SHC).

To be able to estimate every table as accurately as possible, each estimate is based on the largest possible number of records. Tables that contain register variables only are counted from the registers. Tables that contain at least one variable from a survey are estimated from the largest possible combination of registers and surveys.

We guaranteed consistency among the tables by using the technique of repeated weighting. It generates a new set of weights for each estimated table and is based on the repeated application of the regression estimator. When using repeated weighting, the weights of the records in the microdata are adapted in such a way that a new table estimate is consistent with all earlier table estimates.

2.2 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 (fiscal administration) 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 these data files are all 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). In that case multiple 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:

- sex;
- date of birth;
- address[1].

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 not person but address related, which may cause linkage problems if someone has recently moved. When linking the PR and the LFS with this 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 close to 100 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 one 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

[1] In fact, the combination of a postal code (mostly related to the street) and house number is used as substitute for the address. The postal code in the Netherlands consists of four figures, followed by two letters.

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.

Nowadays, the PR is serving as a sampling frame for the LFS. Therefore, the matching rate is almost 100 percent, and no more linkage selectivity problems occur.

2.3 Micro integration

Successfully linking the PR with all the other data sources mentioned, makes much more coherent information on the 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 recognize the strong and weak points of all the data sources used.

Since there are differences between sources, we need a micro integration process to check data and 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 we can use another source. Another advantage of integration is that users of statistical information will get one figure on each social phenomenon, instead of a confusing number of different 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, 2000):

a. harmonisation of statistical units;

b. harmonisation of reference periods;

c. completion of populations (coverage);

d. harmonisation of variables, in case of differences in definition;

e. harmonisation of classifications;

f. adjustment for measurement errors, when corresponding variables still do not have the same value after harmonisation for differences in definitions;

g. imputations in the case of item nonresponse;

h. derivation of (new) variables; creation of variables out of different data sources;

i. checks for overall consistency.

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

Now an example follows 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 though 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.

2.4 The Social Statistical Database (SSD) and repeated weighting

The micro linkage and micro integration process of all the available data sources result in the end in the Social Statistical Database (SSD), a whole set of integrated microdata 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 can be found in Arts and Hoogteijling (2002).

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 as such. Furthermore, 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.

We guaranteed consistency among the census tables by using the technique of repeated weighting. The method of repeated weighting has been described extensively in Houbiers et al. (2003) and Houbiers (2004). It is based on the repeated application of the regression estimator and generates a new set of weights for each table estimated. The results of five

simulation studies testing various aspects of repeated weighting can be found in Van Duin and Snijders (2003). When using repeated weighting, the weights of the records in the microdata are adapted in such a way that a new table estimate is consistent with all earlier table estimates.

To apply the technique of repeated weighting we used the latest version of the software package VRD developed by Statistics Netherlands. The letters VRD stand for Vullen (Filling) Reference Database and the aim of the application is to fill and manage the reference database. The main functions of VRD are the estimating of tables via repeated weighting, adding these tables to the reference database, and the withdrawing aggregates from the reference database. Under the condition of small, independent samples, the variances of the table values can also be estimated. The estimating of the tables does not occur in VRD itself, but takes place in Bascula 4.0 automatically without the VRD-user seeing this explicitly. Estimating the tables and the variances can be done in the batch or interactively.

To be able to estimate every table as accurately as possible, every estimate is based on the largest possible number of records. Tables that contain register variables only, are counted from the registers. Tables that contain at least one variable from a survey are estimated from the largest possible combination of registers and surveys. The combination of registers and surveys form blocks from which the census tables have been estimated. By way of illustration six blocks have been displayed below on basis of which the census tables for the economic active population (employed and unemployed together) were estimated.

1. The register block.
2. The NACE block (all records from the register block for which the international code for economic activity NACE is known, and also the non-employed).
3. The SEE block.
4. The cross-section between NACE and SEE blocks.
5. The Economic Activity block (in fact this is the LFS block, supplemented with information on the employed and retired).
6. The LFS block.

Blocks 2 up to and including 6 were compiled on the basis of survey data. To produce estimates for the complete population, weights have to be determined. These weights depend on:

the precise composition of the block concerned (one or more surveys);
the design of the survey(s);
the non-response correction of the survey(s);
the reduction of the variance by means of auxiliary information;
the reaching of consistency.

Complete consistency is not always possible, for example if too many restrictions were imposed. In some cases complete consistency is possible, but it leads to a very large variation in the weights and thus increases variance drastically. In those cases it is better to restrict the detail that is published.

In compiling the census tables we adapted the weights of the blocks at every VRD turn by means of all relevant register counts and the tables estimated earlier from the blocks. This way, all tables are mutually consistent. Every table has to be calculated from the largest block from which the table can be determined. If all tables are estimated this way with the correct weights, the tables' results are mutually consistent. By starting every time from the largest block, the most detailed possible census tables have been achieved.

The figures of the 2001 census relate to persons living in the Netherlands on 1 January 2001 (counting unit persons). The persons who were living in the Netherlands at the beginning of that day according to the PR were 'counted' in the virtual census. Most of the Dutch population lives in private households, the others are part of institutional households. The number of employees in the tables relates to the end of the year 2000 for which 22 December 2000 was used as reference date to fix the number of jobs of employees in the Netherlands. It was impossible to have a reference day in 2001 for the number of employees since the SSD datasets 2001 were not available on time to use in the 2001 census. The SSD data used registers information on the jobs of employees. If an employee holds several jobs at the same time, he or she can appear several times in the employee register. In the set of tables the features of the main job are used, in which the main job of an employee has been defined as the job with the highest gross wage for the social insurances.

The 2001 census was compiled partly on the basis of sample data. Therefore, margins of inaccuracy have to be taken into account for some results of the 2001 census. Because of the reliability of the results, rules of thumb are being applied for cell values that are based on a sample from the census population. The exact margins of inaccuracy cannot be given because of composing blocks from the surveys and the complex design of these surveys. The rules of thumb have been deduced on the basis of the assumptions that the two LFS datasets (for 2000 and 2001) form one sample and that the 'inclusion probabilities' for this sample were given by the block weights of the LFS block. The rules of thumb for records of observations from the LFS run as follows:

- Table cells based on less than 10 persons are always suppressed.
- Table cells based on 25 or more persons are always published.
- Table cells based on 10–24 persons are only published if they form a part of a breakdown (by age or sex), in which no cells based on less than 10 persons occur, and at least 50 percent of the cells in the breakdown have more than 25 persons. The threshold of 25 persons corresponds to an estimated relative inaccuracy of at most 20 percent (i.e. the estimated margins amount to 40 percent at most).

The rules of thumb for records from the SHC are of the same form. However, somewhat higher threshold values are applied because of the fact that the sample size of the SHC is somewhat more limited than the one of the LFS. For table cells with households or dwellings as counting unit, analogous rules of thumb are applied for the Dutch census.

3. KEY RESULTS OF THE 2001 CENSUS IN THE NETHERLANDS

3.1 Population by sex, age and type of household

At the start of 2001 a total of 16.0 million people were living in the Netherlands, 7.9 million male and 8.1 million female. In the age categories 0-14 and 15-74 year there were some more males than females, but in the category 75 year and older there were almost twice as many women than men. Most people live in private households. More than 200 thousand people lived in institutional households, such as health care institutions and institutions for retired and elderly people. About 36 percent of this group was male and 64 percent female. Of the people in institutional households 57 percent was over 75. This group is dominated by women. More information about the population by sex, age and type of household can be found in Table 1.

Table 1. Population by sex, age and type of household

Sex and type of household		Age in years		
		0-14	15-74	75+
Total population	15,985,538	2,977,283	12,036,171	972,084
Male	7,909,052	1,522,811	6,047,425	338,816
Female	8,076,486	1,454,472	5,988,746	633,268
		0-14	15-74	75+
Population in private households	15,766,606	2,970,545	11,947,996	848,065
Male	7,829,914	1,518,611	5,998,189	313,114
Female	7,936,692	1,451,934	5,949,807	534,951
		0-14	15-74	75+
Population in institutional households	218,932	6,738	88,175	124,019
Male	79,138	4,200	49,236	25,702
Female	139,794	2,538	38,939	98,317

3.2 Population by economic activity

At the start of 2001 just under half of the people living in the Netherlands belonged to the economically active population (labour force). The working labour force included 7.4 million people: 6.8 million are employees and 0.6 million self-employed. The unemployed labour force comprised almost 200 thousand people. In the organisational set-up of the census,

employees, the self-employed and unemployed are mutually exclusive categories. Self-employed people who also work a number of hours a week for pay are counted as employees. Someone in the working labour force cannot be unemployed at the same time. The number of unemployed is estimated on the basis of sample information.

Of the economically active population 58 percent was male, while of the economically inactive population 58 percent was female. The economically inactive include attendants at educational institutions, retired people and people engaged in family duties. The number of housewives is more than 18 times the number of househusbands. More information about the population by sex, age and type of household can be found in Table 2.

Table 2. Population by economic activity

Population by economic activity		***Male***	***Female***
Economic active population	**7,586,914**	**4,388,239**	**3,198,675**
Working	**7,394,777**	**4,287,967**	**3,106,810**
Employed	**6,786,511**	**3,883,813**	**2,902,698**
Self-employed	**608,266**	**404,154**	**204,112**
Unemployed	**192,137**	**100,272**	**91,865**
Economic inactive population		***Male***	***Female***
All ages	**8,398,624**	**3,520,813**	**4,877,811**
15-74	**4,449,257**	**1,659,186**	**2,790,071**
Attendant at educational institutions	**640,446**	**342,934**	**297,512**
Retired	**1,355,940**	**620,493**	**735,447**
Engaged in family duties	**1,270,420**	**65,821**	**1,204,599**
Other economically inactive	**1,182,451**	**629,938**	**552,513**

Working population by branch of economic activity

The 7.4 million members of the working population can be divided by branch of economic activity by means of the NACE code. For an employee who has more than one job we took the features of his or her main job. In the context of the Dutch census, the main job of a person has been defined as the job that yielded the highest wage for the social insurances in 2000. Counted this way, the Netherlands had 0.2 million people working in agriculture and fishing, 1.5 million in manufacturing and construction and 5.7 million in services at the start of 2001. Of those working in services, 3.5 million worked in commercial services and over 2.1 million in non-commercial services.

Employees by working hours

An interesting phenomenon is how many hours a week employees work in their main job. Of the almost 6.8 million employees in the Netherlands 4.2 million employees work full-time (≥35 hours a week), 1.8 million employees have a long part-time job (less than 35 hours, but at least 15 hours a week) and 0.8 million have a short part-time job (less than 15 hours per

week). Of those working full-time 77 percent is male, while of the part-timers 75 percent is female. More information about the working hours of employees can be found in Table 3.

Table 3. Employees by working hours

Employees by working hours		Male	Female
Employees	6,786,511	3,883,813	2,902,698
Full-time (≥35 hours a week)	4,222,228	3,236,504	985,724
Part-time total	2,564,283	647,309	1,916,974
Long part-time (15-<35 hours a week)	1,793,656	419,071	1,374,585
Short part-time (<15 hours a week)	770,627	228,238	542,389

3.3 Working population by occupation

By means of the International Standard Classification of Occupations (ISCO) working people can be classified by occupation. For men the most common occupation categories in 2001 were:

professionals;
legislators, senior officials and managers;
craft and related trades workers.

For women the occupation categories were:

technicians and associate professionals;
clerks;
service workers and shop and market sales workers.

More information about the working population by occupation can be found in Table 4.

Table 4. Working population by occupation

Working population by occupation			Male	Female
Working population		7,394,777	4,287,967	3,106,810
1	legislators, senior officials and managers	926,631	695,563	231,068
2	professionals	1,205,163	705,357	499,805
3	technicians and associate professionals	1,248,759	607,819	640,939
4	clerks	841,219	271,862	569,358
5	service workers and shop and market sales workers	800,629	259,173	541,456
6	skill agricultural and fishery workers	105,256	78,280	26,976
7	craft and related trades workers	712,093	677,256	34,837
8	plant and machine operators and assemblers	446,722	398,845	47,877
9	elementary occupations	522,901	272,435	250,467
0	armed forces	37,032	34,227	2,805
99	occupation unknown	548,374	287,151	261,223

3.4 Population by level of education

The population living in the Netherlands can be classified by level of education by means of the International Standard Classification of Education (ISCED). Actually, it is the highest level of educational attainment that determines the category by which a person is classified in the ISCED. Of the 12.0 million people aged between 15 and 75 years the most common level of education is the secondary level. The number of people with a tertiary level of education is larger than the number of people with a primary level of education. For the group aged over 75 the secondary level is also the most common, but there are also considerably more people with primary than with a tertiary level of education. More information about the population by level of education can be found in Table 5.

Table 5.Population by level of education

Population by level of education		***0-14***	***15-74***	***75+***
Total population	**15,985,538**	**2,977,283**	**12,036,171**	**972,084**
No education at all	**1,244,031**	**1,244,031**	**0**	**0**
Pre-primary education (ISCED 0)	**1,370,511**	**1,198,580**	**154,832**	**17,098**
Primary education (ISCED 1)	**2,787,104**	**534,672**	**1,825,655**	**426,778**
Lower secondary education (ISCED 2)	**3,145,529**		**2,924,405**	**221,125**
Upper secondary education (ISCED 3c)	**2,711,384**		**2,566,372**	**145,012**
Upper secondary education (ISCED 3b)				
Upper secondary education (ISCED 3a)	**1,873,656**		**1,828,072**	**45,584**
Post secondary non-tertiary education (ISCED4)	**483,684**		**468,699**	**14,985**
First stage of tertiary education (ISCED5b)	**247,194**		**238,029**	**9,165**
First stage of tertiary education (ISCED5a)	**2,081,590**		**1,992,670**	**88,920**
Second stage of tertiary education (ISCED6)	**32,760**		**31,082**	**1,678**
Education unknown	**8,094**	**0**	**6,356**	**1,738**

4. THE 2001 CENSUS COMPARED TO EARLIER DUTCH CENSUSES

The first census in the Netherlands was held in 1795 for the purpose of establishing voting constituencies. At that time the united provinces of the Netherlands were still a republic and the borders were different from the current borders. After Napoleon the Netherlands became a kingdom and once every ten years a census was held. The first census in the Kingdom of the Netherlands was held in 1829. Before Statistics Netherlands was established, another six censuses were held in 1839, 1849, 1859, 1869, 1879 and 1889 under the responsibility of the Ministry of the Interior. In 1899 Statistics Netherlands was established and was put directly in charge of the eighth census. In the 20th century six more traditional censuses were carried out in 1909, 1920, 1930, 1947, 1960 and 1971. The three most recent censuses (1981, 1991 and 2001) were not based on a complete enumeration but on registers and surveys available for Statistics Netherlands.

Originally, the censuses had two aims. First, they were meant to correct errors in the municipal population registers. Second, they were used to get extra information about the socio-economic phenomena in the country. Since the Netherlands conducts a register-based census, the first aim no longer exists. Also, the quality of the central Population Register (PR), which unites all municipality population registers, has improved considerably over time. This is because the incentive for municipalities to keep their population registers up-to-date is the allocation of central government funds among municipalities. This is mainly based on the population size according to the local registers. Another reason is that it is extremely difficult to function in Dutch society without being included in the PR. So both municipalities and citizens have enough incentives to keep the PR of good quality. The second aim is still valid and many census results are published in a historical or international context. Currently, census data are popular for comparisons between countries.

Table 6 presents some key results of the Dutch censuses in the period 1829-2001. Remarkable is the ageing of the Dutch population, especially in the post-war period.

Table 6. Population by age category in the period 1829-2001

Number of the census	Year of the census	Total population	0-19	20-64	65+
		× 1,000	in % of the total population		
1	1829	2,613.3	44	50	5
2	1839	2,860.6	45	50	5
3	1849	3,056.9	43	53	5
4	1859	3,309.1	42	53	5
5	1869	3,579.5	43	52	6
6	1879	4,012.7	44	50	5
7	1889	4,511.4	45	49	6
8	1899	5,104.1	44	50	6
9	1909	5,858.2	44	50	6
10	1920	6,865.3	42	52	6
11	1930	7,935.6	40	54	6
12	1947	9,625.5	38	55	7
13	1960	11,462.0	39	53	9
14	1971	13,060.1	36	54	10
15	1981	14,216.9	31	57	12
16	1991	15,070.0	25	62	13
17	2001	15,985.5	24	62	14

5. THE DUTCH 2001 CENSUS COMPARED TO OTHER COUNTRIES

More than fifty countries participated in the 2001 Census Round. Most countries chose a day in 2001 as their reference day, but unfortunately they chose many different days. As it will take a long time before all countries finish the tables required by the international organisations, the Netherlands took the initiative to make some simple comparisons among eight countries that were relatively quick in compiling the set of tables for Eurostat. The nine countries are the Netherlands (NL), Norway (NO), Sweden (SE), Finland (FI), Estonia (EE), Switzerland (CH), Slovenia (SI), Greece (GR) and the United Kingdom (UK). The nine countries differ in size, but all except the United Kingdom have a fairly limited number of inhabitants compared to France and Germany.

The nine countries are members of the European Union (EU) or the European Free Trade Association (EFTA). The Netherlands joined the European Community at the start in 1958, the United Kingdom joined in 1973 and Greece in 1981. The European Community became the European Union in 1995 when Sweden and Finland joined. Estonia, the most northern of the ten accession countries, and Slovenia, the most northern part of the former Yugoslavia, joined the EU in 2004. Norway and Switzerland are EFTA members and work closely together with the EU countries. Norway is also a member of the European Economic Area (EEA). The EEA agreement came into force on 1 January 1994. EEA countries are the EU 15, Norway, Iceland and Liechtenstein. Switzerland did not join the EEA, but works together with the EU countries on a bilateral basis. Statistics is one of the issues on which the EEA countries work together. The aim of the statistical co-operation in the EEA is to build a European Statistical System that gives a coherent and comparable description of the economic, social and environmental developments in the EEA countries.

The nine countries that are compared have different reference dates: 31 March 2000 (Estonia), 5 December 2000 (Switzerland), 1 January 2001 (The Netherlands, Sweden and Finland), 18 March 2001 (Greece), 29 April 2001 (United Kingdom), 3 November 2001 (Norway) and 31 March 2002 (Slovenia).

Table 7 presents the estimated costs of the 2001 censuses, and the population and the area of the nine countries. Estonia, Slovenia, Greece and the United Kingdom held a traditional census, Switzerland used a combination of a traditional census and register information to produce the census tables. Norway relied largely on registers but conducted a census for some missing housing variables. Sweden and Finland held an entirely register-based census and the Netherlands has a virtual census based on existing registers and surveys. The 2001 census costs for Norway, Estonia, Switzerland, Slovenia, Greece and the United Kingdom include enumeration costs. In the Netherlands, Sweden and Finland such enumeration costs do not

exist for the 2001 census, so the costs presented in Table 7 for these three countries are rough indicators of the extra costs of producing census tables for the international organisations and of analysing and publishing the results. Table 7 shows that the costs per inhabitant in those countries that held interviews for the census were much higher than the countries that did not have enumeration costs. In Table 7 the population densities among the nine countries can be compared. The Netherlands has the highest population density, followed by the United Kingdom and Switzerland. The population density in the Nordic countries (Norway, Sweden and Finland) and in Estonia is relatively low. Slovenia and Greece take up a middle position.

Table 7. Comparison of nine countries according to the 2001 census results

Country	NL	NO	SE	FI	EE	CH	SI	GR	UK
Cost of the census (in millions of Euros)	3.0	14.6	1.0	0.8	10.2	99.1	8.0	49.7	367.4
Population (× 1,000,000)	16.0	4.5	8.9	5.2	1.4	7.3	2.0	10.9	58.8
Area (× 1,000 km^2)	41.5	323.9	450.0	338.1	45.1	41.3	20.3	132.0	244.1
Cost of the census per inhabitant (in Euros)	0.2	3.2	0.1	0.2	7.3	13.6	4.0	4.6	6.2
Population density (persons per km^2)	386	14	20	15	31	177	99	83	241

Table 8 presents some simple demographic comparisons of 2001 census data on the nine countries. The data presented in Table 8 are calculated from the tables produced for the international organisations and sent to Eurostat. They were checked by the different countries that produced the tables. Unfortunately, the indicator on the percentage of non-nationals in Table 8 could not be calculated for the United Kingdom since some relevant information is missing in the British census tables. However, interesting analyses have been conducted to compare the 2001 census results in the Netherlands and the United Kingdom. These analyses can be found in section 6. Remarkable differences exist between the nine countries. The percentage of women is by far the highest in Estonia. The percentage of singles is high in the Nordic countries, low in Switzerland and especially low in Greece. For the percentage of singles the Netherlands, Estonia, Slovenia and the United Kingdom take up a middle position. The percentage of non-nationals is high in Estonia and Switzerland, where it is relatively difficult to get the nationality. Finland and Slovenia have an extremely low percentage of non-nationals. In Slovenia it includes only population with stated foreign citizenship. Finland also has an extremely low percentage of people born outside the country. Most countries have much higher percentages of people born abroad than percentages of non-nationals. However, we have to realise that not all non-nationals are persons born outside the country. In other words, these groups are not nested.

Table 8. A demographic comparison according to the 2001 census results

Country	NL	NO	SE	FI	EE	CH	SI	GR	UK
Percentage of women	50.5	50.4	50.5	51.2	53.9	51.0	51.2	50.5	51.4
Percentage of singles	44.7	48.4	49.8	47.1	44.1	42.1	44.5	39.7	44.3
Percentage of non-nationals	4.2	4.1	5.4	1.8	20.0	20.5	1.9	7.0	n.a.
Percentage of people born outside the country	10.1	6.9	11.3	2.6	19.2	21.6	8.6	10.3	8.3

Table 9 presents two indicators to make a simple economic comparison of the nine countries. Different definitions and ways of collecting the data may hamper the comparisons. The percentage of economically inactive population is relatively high in Greece. For this indicator Switzerland has the lowest score. When we compare the percentages of unemployed distinct groups can be discerned. Finland, Estonia and Greece have many unemployed people, whereas in the Netherlands, Norway, Sweden and Switzerland unemployment was low. Slovenia and the United Kingdom take up a middle position.

Table 9. An economic comparison according to the 2001 census results

Country	NL	NO	SE	FI	EE	CH	SI	GR	UK
Percentage of economically inactive population	52.5	48.5	51.8	50.8	53.3	45.8	51.7	57.8	52.1
Unemployed as percentage of the economically active population	2.5	2.3	3.8	12.5	13.8	4.0	6.7	11.0	5.7

Table 10 presents two educational indicators to compare the nine countries. Different definitions and ways of collecting the data may hamper the comparisons as well in Table 10. All countries have difficulties fitting their national education classification into the ISCED codes. For all countries compared except Estonia and Slovenia the OECD (2003) study also gives some information about the education indicators in 2001. A somewhat different population is considered and therefore the absolute results are hard to compare. However, the relative differences in percentages of people with primary and tertiary education in the OECD study agree to a reasonable extent to the results of the censuses. Norway has a high quality education register and an extremely low percentage of the population aged 15-74 with a primary level of education or less. Also Slovenia has a very low percentage of primary educated people. Finland, Switzerland, Greece and the United Kingdom have a rather high percentage of people with primary education. The percentage of the population aged 15-74 with a tertiary level of education differs less among the nine countries than those with a

primary level. Estonia, Finland and Norway have the highest percentages and the United Kingdom, Greece and Slovenia the lowest.

Table 10. An educational comparison according to the 2001 census results

Country	***NL***	***NO***	***SE***	***FI***	***EE***	***CH***	***SI***	***GR***	***UK***
Percentage of the population 15-74 with a primary level of education or less[a)]	**16.5**	**4.1**	**15.6**	**37.2**	**11.3**	**33.5**	**5.5**	**38.1**	**37.0**
Percentage of the population 15-74 with a tertiary level of education	**18.8**	**22.9**	**20.2**	**24.6**	**24.8**	**17.5**	**13.3**	**14.4**	**14.6**

[a)]: including level of education unknown

6. COMPARISON OF THE UK AND NETHERLANDS CENSUS DATA

6.1 Population structure

Figure 1 and Table 11 show that the Dutch population is younger than the UK population in age structure. The Netherlands has more of its population distributed in working ages and under age 5, and less over age 60 than the UK. Both the mean and the median ages of the populations reflect this. The mean age of the UK population is 38.6 years and the mean age of the Dutch population is 37.8 years. The median ages for the populations are 37.9 for the UK and 37.5 the Netherlands.

Table 11. The population in the United Kingdom and the Netherlands by age and sex (in %)

Age group	Males		Females		Total	
	UK	NL	UK	NL	UK	NL
0-15	21,25	20,47	19,15	19,15	20,17	19,80
16-59	60,13	63,58	58,01	60,36	59,04	61,95
60+	18,62	15,96	22,84	20,49	20,79	18,25
0-15	21,25	20,47	19,15	19,15	20,17	19,80
16-79	75,97	77,53	75,41	76,41	75,68	76,97
80+	2,77	2,00	5,45	4,44	4,15	3,23

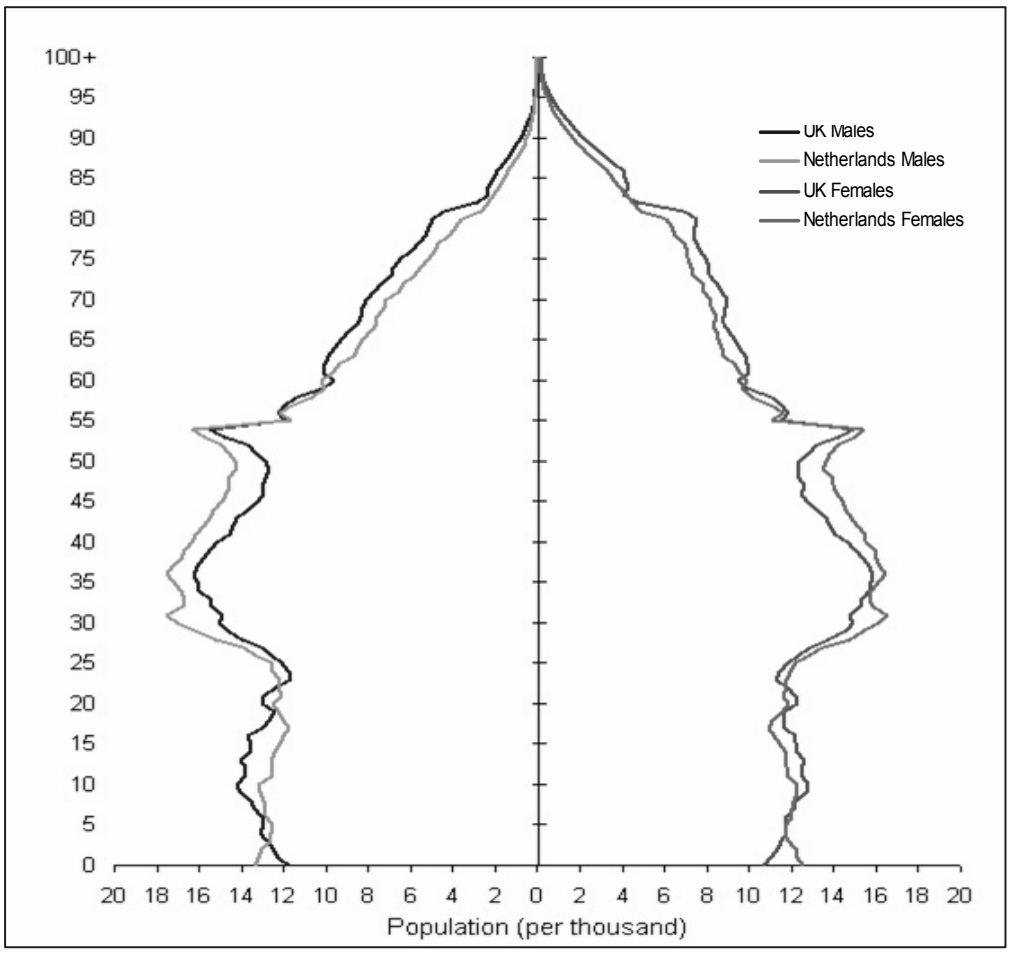

Figure 1. The population in the United Kingdom and the Netherlands by age and sex

Figure 1 shows a standardised population pyramid for the UK and the Netherlands. The population structures are similar in each country for both men and women. However, there are several differences between the UK and the Netherlands. Under age 5 there are proportionately fewer people in the UK population compared to the Netherlands and then over age 5 the UK has relatively more people until age 22. The Netherlands has a higher proportion of its population between ages 22 and 55 than the UK. Then at older ages, in particular over 60 years old, the UK has relatively more people than the Netherlands.

Table 11 shows differences in the percentage population distributions. The UK has around 2.5 percent fewer of its population aged 16-59 years, also the Netherlands has a smaller percentage of its population that are very elderly (over 80 years old).

The dependency ratios (dependants defined as aged 0-15 or 65 and over) in Table 12 shows that the UK has a higher population dependency ratio than the Netherlands. The pattern of dependency is different between the UK and the Netherlands; in the Netherlands children are a bigger component of the dependants ratio than the elderly, whereas the opposite is true for the UK.

Table 12. Dependency ratios in the United Kingdom and the Netherlands

	Youth	Aged	Overall
	(0-15)	(65+)	(0-15 & 65+)
UK	0,34	0,35	0,69
NL	0,32	0,29	0,61

The sex ratios in Figure 2 show the Netherlands does not lose their excess of males until age 60-64, whereas the UK sex ratio falls below 1 at age 20-24. This is likely to be the reflection of different patterns of migration by sex for the United Kingdom and the Netherlands. At older ages there is a slightly greater excess of females in the Netherlands population compared to the UK population.

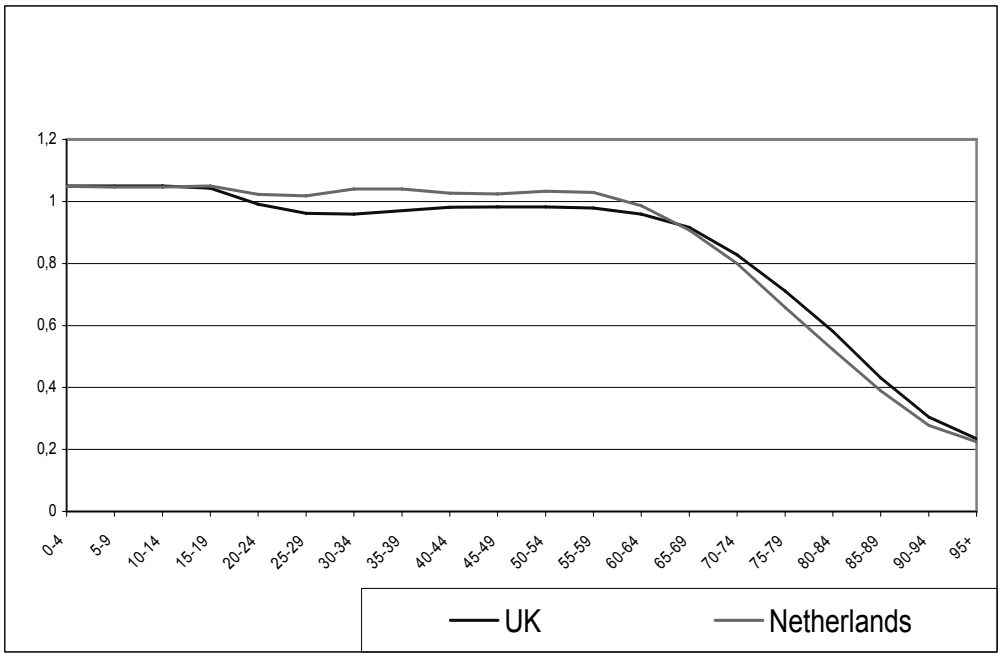

Figure 2. Sex ratios in the United Kingdom and the Netherlands at five-year age intervals

6.2 Population by marital status

The patterns of marital status by five-year age groups shown in Figure 3 and Figure 4 are similar between the male populations of the Netherlands and the UK. There are some small differences, such as slightly higher levels of married men in the UK under age 35, in particular at age 25-29 where compared with the Netherlands 3.5 per cent more of the male population are married in the UK. There are also somewhat higher proportions of divorced in younger and middle age groups in the UK, although the UK and the Netherlands display similar levels of proportions divorced at older ages (over 65 years). The male population of

the Netherlands has relatively more men who are married in every age category between age 35 and 89.

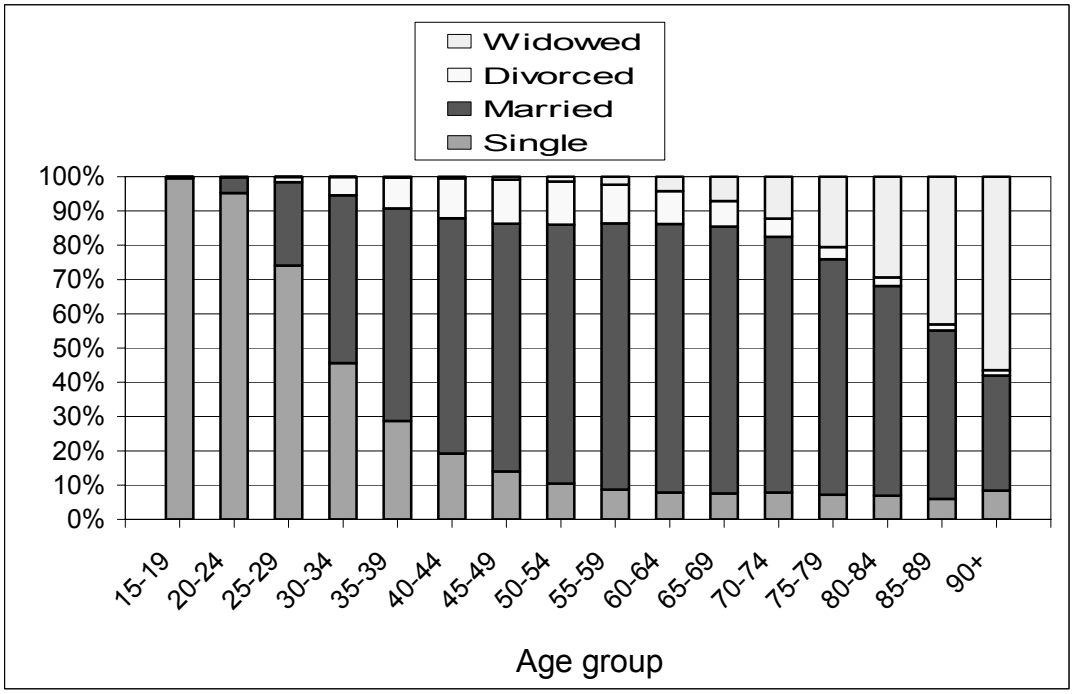

Figure 3. United Kingdom male population by marital status and age group

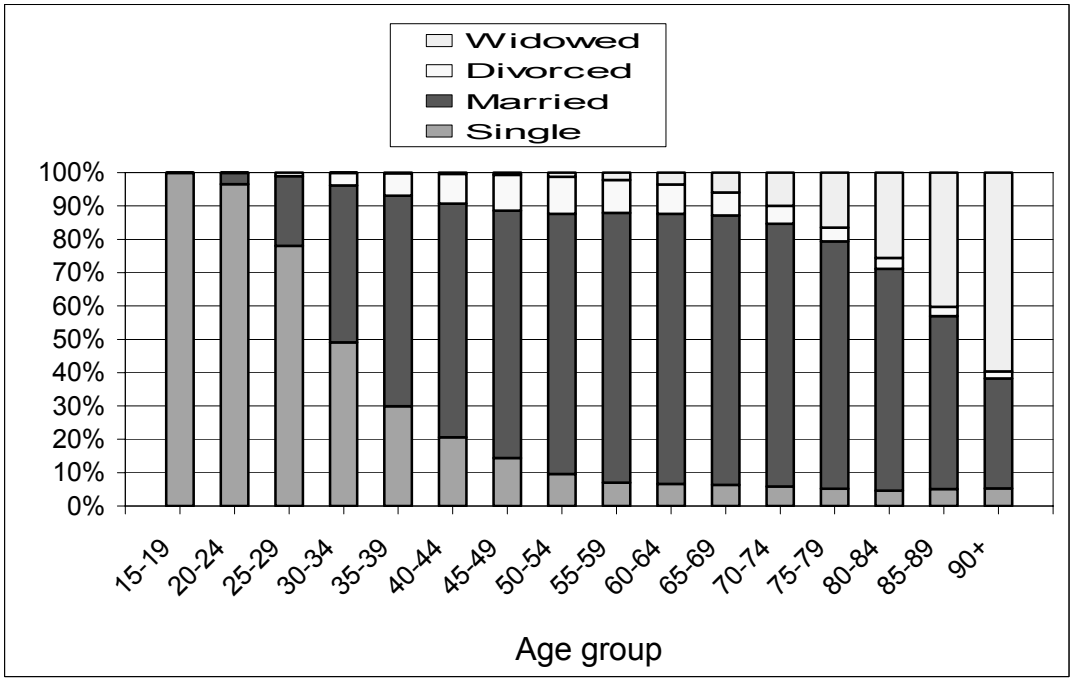

Figure 4. Netherlands male population by marital status and age group

Figures 5 and 6 show that marital status patterns are similar between the female populations of the UK and the Netherlands, in particular in the youngest three age groups and over age 65. Compared with the UK, there are proportionately more women married in the Netherlands in every age group from age 30 to 74. In particular between ages 30-44 3 percent more women

are married in the Netherlands compared to the UK. Below age 65 the percentage of the female population classified as divorced is greater in the UK, as with the male population. Between ages 35 and 49 there are 3 percent more divorced women in the UK population than in the Dutch population.

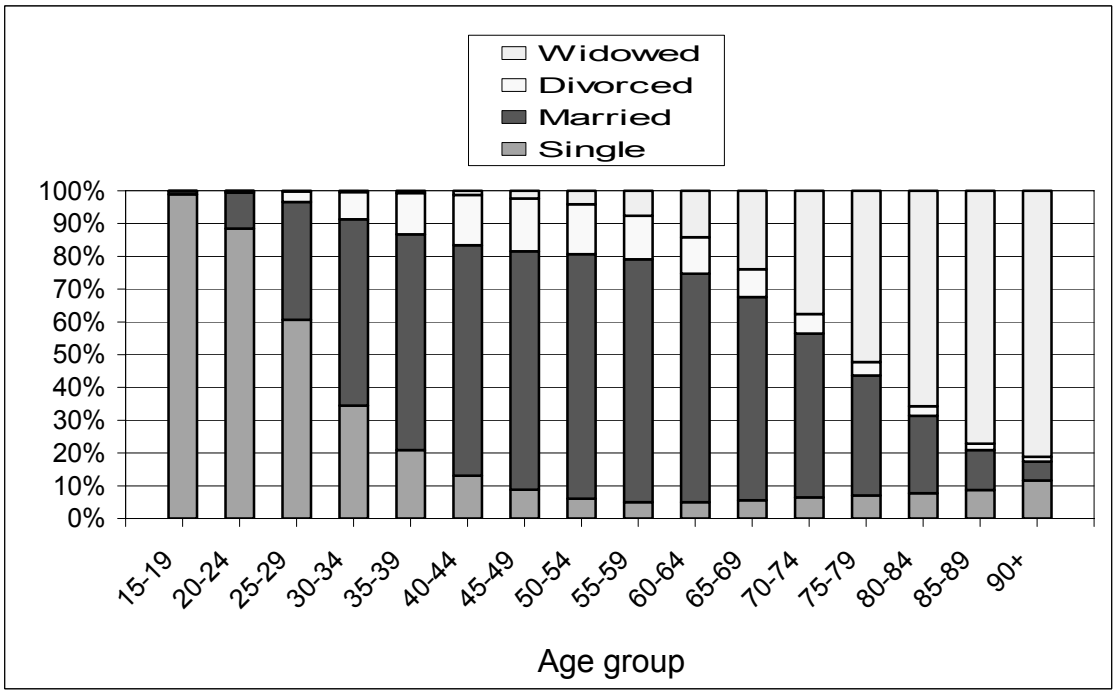

Figure 5. United Kingdom female population by marital status and age group

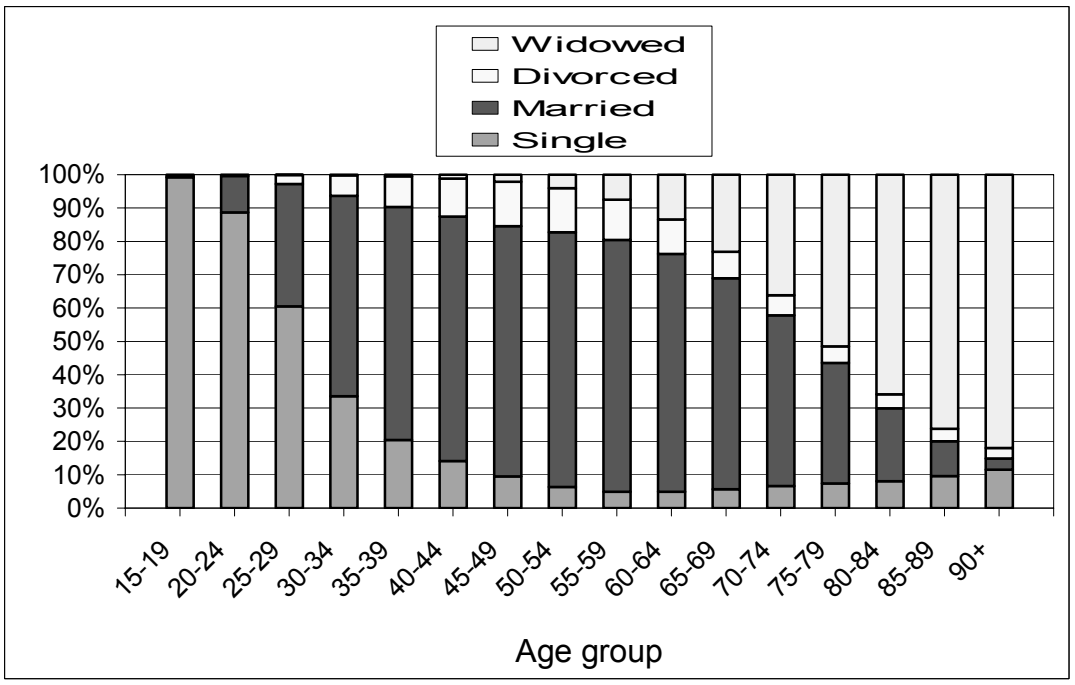

Figure 6. Netherlands female population by marital status and age group

6.3 Population by household type[2]

A comparison of population by household type has been provided at Table 13. However, it is not apparent from the supporting information for the UK and the Netherlands tables whether there are definitional differences in household classification between the two countries, in particular with the child subcategory. Child here appears to mean in parent's household, so is not bound by age.

Table 13. Adult population in the United Kingdom and the Netherlands by household type and sex (in %)

		Private Household[a)]					Institution
		Child	Spouse	Cohabitant	Lone parent	Living alone	
Males	UK	34,2	42,8	8,1	1,4	11,7	1,8
	NL	31,8	44,3	8,7	0,8	13,4	1,0
Females	UK	27,6	40,3	7,5	8,1	14,6	2,0
	NL	26,5	43,3	8,4	4,2	15,9	1,8
Total	UK	30,8	41,5	7,8	4,8	13,2	1,9
	NL	29,1	43,8	8,6	2,5	14,7	1,4

[a)]: the other subcategory within the private household category was excluded from the analysis for both the UK and the Netherlands

Table 14 shows that for the elderly population aged 70 and over, a much higher proportion of the elderly in the Netherlands live in health care institutions and institutions for retired or elderly persons.

Table 14. Elderly population in institutions[a)] in the United Kingdom and the Netherlands by sex (in %)

		Age group		
		70-79	80-89	90+
Males	UK	1,5	5,5	18,2
	NL	1,9	10,2	33,5
Females	UK	2,1	10,3	31,7
	NL	3,3	18,4	47,1
Total	UK	1,8	8,7	28,7
	NL	2,7	15,7	44,2

[a)]: institutions are defined as health care institutions and institutions for retired or elderly persons

[2] All data in this section are analysed on a person rather than on a household basis. Both the Dutch, but in particular the UK census data contain a few occurrences of improbable situations, such as children aged under 10 years old classified as living with a spouse and persons over 90 years old classified as child. However, the numbers that are classified in this way are small and do not effect the percentage distributions to one decimal place.

6.4 Foreign born population

Table 15 shows the foreign born population distributions of the Netherlands and the UK are broadly similar. In both the UK and the Netherlands the foreign born population make up around one in ten of the total population.

Table 15. Population by region of birth in the United Kingdom and the Netherlands (in %)

	Region of birth								
	Parent Country	Europe	Middle East	Asia	North America	Central & South America	Africa	Oceania	Other
UK	91,67	2,89	0,17	2,51	0,40	0,57	1,42	0,29	0,07
NL	89,90	3,81	0,32	2,01	0,19	1,44	1,72	0,08	0,53

The Netherlands has 1.5 per cent more people in its population who were born outside the country than the UK. The regions with noticeably different percentages between the UK and the Netherlands are Europe and Central and South America. The Netherlands has a much higher percentage of residents born in South or Central America than the UK. Conversely, the UK has higher percentages for Asian and North American born.

Tables 16 and 17 show the age distributions of the foreign born populations are similar in the Netherlands and in the UK. The smallest percentage of the population not born in the country is in ages 0-14 in both the UK and the Netherlands. Then, compared with the UK, the Netherlands has a lower percentage of its population born in the country at every age group up to 65 years or older.

Table 16. Population by region of birth and age group in the United Kingdom (in %)

	Region of birth									
Age group	Parent Country	Europe	Middle East	Asia	North America	Central & South America	Africa	Oceania	Other	NL[a)]
0-14	96,71	1,07	0,14	0,78	0,34	0,14	0,64	0,14	0,03	0,05
15-24	91,60	2,93	0,28	2,53	0,45	0,32	1,52	0,32	0,06	0,07
25-34	88,12	3,68	0,25	3,79	0,55	0,44	2,29	0,81	0,07	0,10
35-44	89,37	2,71	0,23	3,63	0,53	0,72	2,39	0,34	0,09	0,08
45-54	90,32	2,94	0,16	3,22	0,39	0,84	1,84	0,20	0,09	0,07
55-64	91,32	3,52	0,10	2,54	0,22	1,04	1,03	0,13	0,10	0,06
65+	92,43	4,02	0,06	1,75	0,29	0,75	0,48	0,13	0,08	0,06

[a)]: these populations are also included within the European population

Table 17. Population by region of birth and age group in the Netherlands (in %)

	Region of birth									
Age group	**Parent Country**	**Europe**	**Middle East**	**Asia**	**North America**	**Central & South America**	**Africa**	**Oceania**	**Other**	**UK[a)]**
0-14	**95,51**	**1,62**	**0,35**	**0,80**	**0,16**	**0,47**	**0,68**	**0,04**	**0,37**	**0,15**
15-24	**88,93**	**3,64**	**0,47**	**1,76**	**0,19**	**1,37**	**2,70**	**0,07**	**0,86**	**0,21**
25-34	**84,38**	**6,32**	**0,58**	**1,78**	**0,28**	**2,35**	**3,42**	**0,15**	**0,76**	**0,42**
35-44	**86,35**	**4,92**	**0,41**	**2,05**	**0,33**	**2,46**	**2,55**	**0,21**	**0,71**	**0,42**
45-54	**89,42**	**3,70**	**0,18**	**2,91**	**0,18**	**1,73**	**1,27**	**0,07**	**0,55**	**0,40**
55-64	**90,53**	**4,20**	**0,10**	**2,44**	**0,08**	**1,11**	**1,19**	**0,02**	**0,35**	**0,26**
65+	**93,46**	**2,67**	**0,03**	**2,83**	**0,04**	**0,58**	**0,27**	**0,01**	**0,11**	**0,12**

[a)]: these populations are also included within the European population

The UK has a higher percentage of residents born in Asia at every age except over 65 than the Netherlands. The percentage of the population born in Europe is similar to the percentage born in Asia for the UK (except over age 65), whereas within the Netherlands the European born population is always the largest of the foreign born groups.

At every age the percentage of people born in the UK living in the Netherlands is higher than the percentage of people born in the Netherlands living in the UK. However, this reflects the relative size of the UK and Dutch populations. Table 18 shows the relative risks for being born in one country and living in the other. This indicates a relationship opposite to the one shown in Tables 16 and 17. At every age there is a much higher 'risk' of being Dutch born and living in the UK than there is of being UK born and living in the Netherlands.

Table 18. Relative risks for being born in one country and living in the other

	Approximate percentage risk of being Dutch born and living in the UK[a]	**Approximate percentage risk of UK born and living in the Netherlands[b]**
0-14	**0.19**	**0.04**
15-24	**0.29**	**0.06**
25-34	**0.40**	**0.14**
35-44	**0.31**	**0.14**
45-54	**0.27**	**0.13**
55-64	**0.26**	**0.08**
65+	**0.28**	**0.03**

[a)]: calculated by dividing the Dutch born population in the UK by the overall Dutch population for each age group

[b)]: calculated by dividing the UK born population in the Netherlands by the overall UK population for each age group

6.5 Population by educational attainment[3]

For Figures 7-10 the categories of lower secondary and upper secondary educational attainment were combined to make an overall secondary educational attainment category. This is because there were large differences between the UK and the Netherlands in the secondary level attainment data, which are likely to be due to differences in classification.

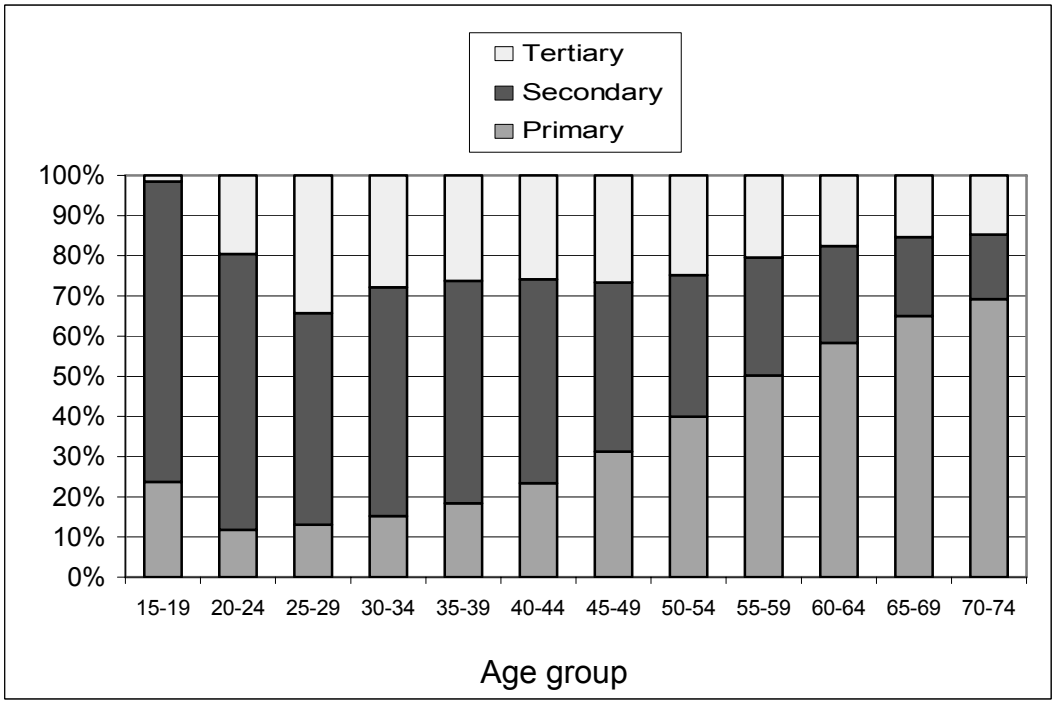

Figure 7. Highest level of educational attainment for the male population in the United Kingdom by age group

[3] There are some issues with the comparability of the education data as overall 9.7 percent of males and 5.9 percent of females in the UK had an unknown educational status. It is likely that this group of people is biased towards lower levels of educational attainment than the population as whole. This could significantly affect the UK's distribution of education levels. Also the Netherlands had a post secondary category that was included with tertiary qualifications, which may have affected the comparison with the UK.

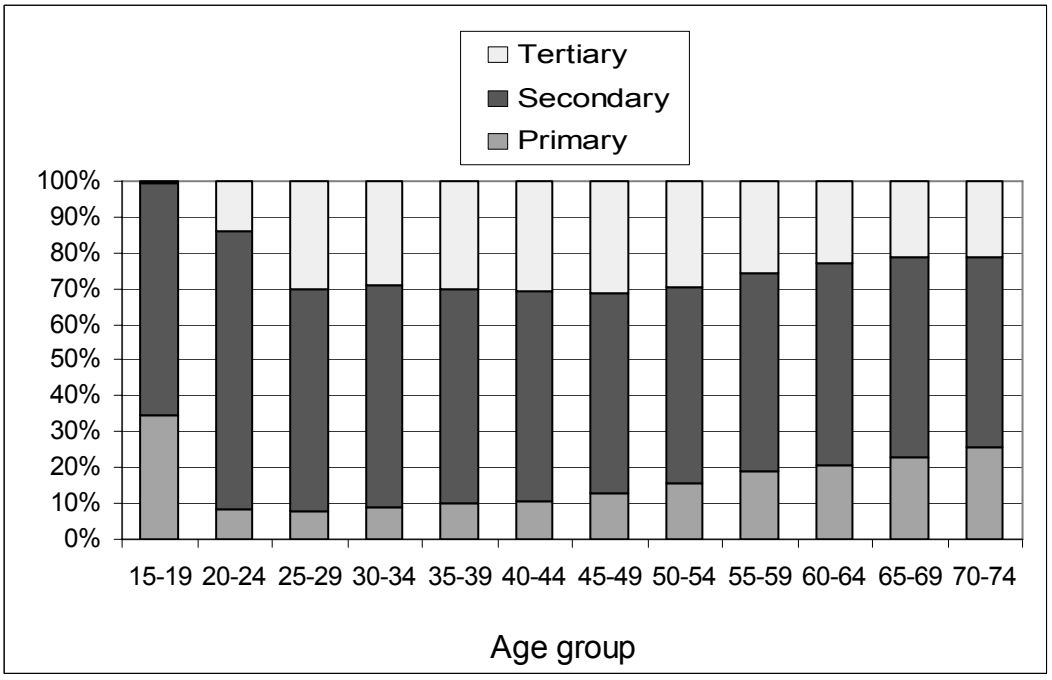

Figure 8. Highest level of educational attainment for the male population in the Netherlands by age group

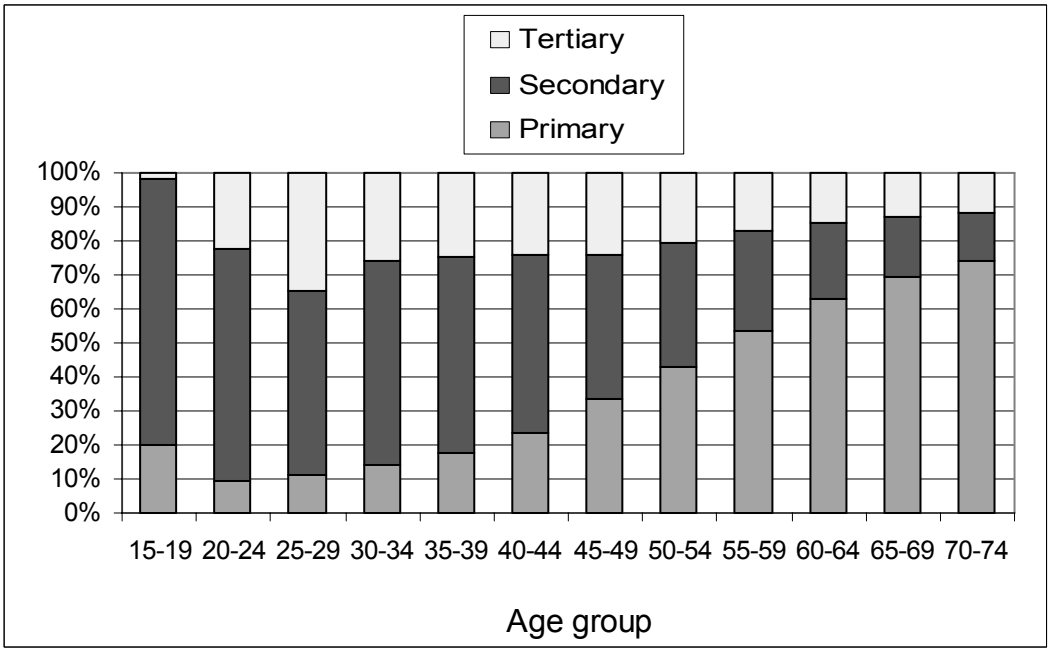

Figure 9. Highest level of educational attainment for the female population in the United Kingdom by age group

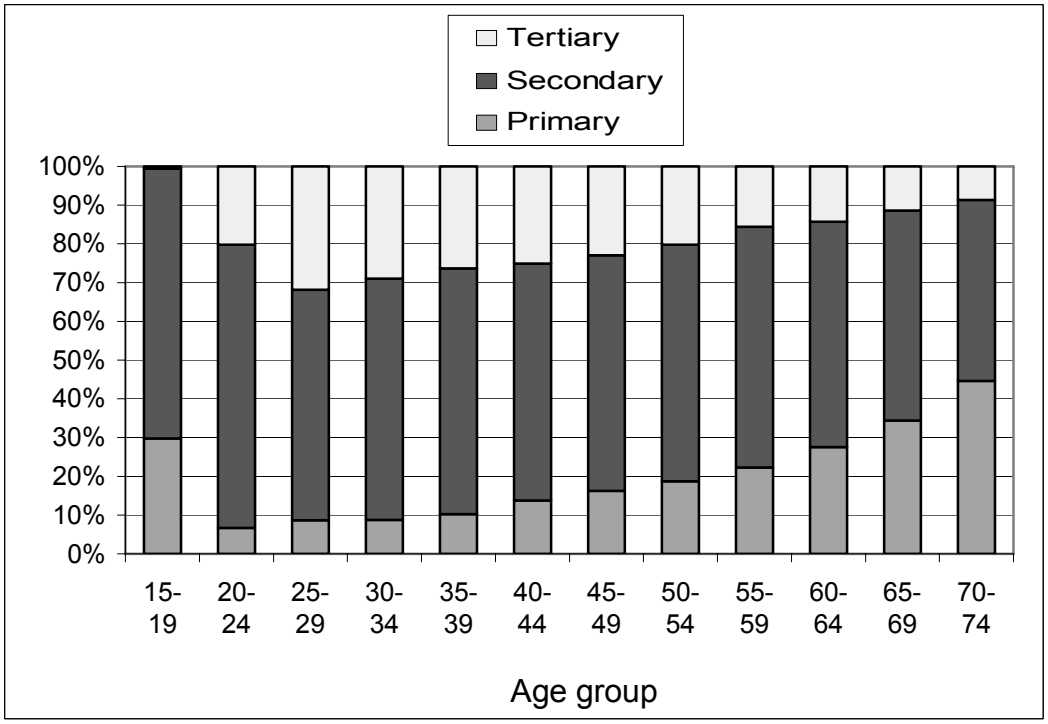

Figure 10. Highest level of educational attainment for the female population in the Netherlands by age group

The patterns of educational attainment shown in Figures 7-10 are quite different between the populations of the Netherlands and the UK. The biggest differences between the UK and the Netherlands are found in the distributions of primary and secondary educational attainment.

The percentage of males and females with only primary level education increases with every age group from 20-24 to 70-74 years old, but the increases are much more marked in the UK populations than in the Dutch populations. In the UK the percentage of people with only a primary education rises from around 10 percent at age 20-24 years to close to 70 percent at age 70-74. Whereas in the Netherlands this percentage rises from roughly 7 percent to 26 percent for men and 45 percent for women.

The pattern of secondary level educational attainment is also different between the UK and the Netherlands. At older ages the percentage of UK population with secondary education decreased quite markedly, as the percentage of the population with only primary education rises. However, relative to the UK the percentage of the Dutch population with secondary education remains quite stable and high over the age range. So above age 55 over 25 percent more of the Dutch population has attained secondary education compared to the UK population. At younger ages the difference is under 10 percent.

Within the UK patterns of educational attainment by age are similar between the sexes. However, within the Netherlands there are quite large gender differences at the older ages.

The distribution of men and women with a tertiary education is broadly similar for the Netherlands and the UK. Younger cohorts generally show higher proportions attaining tertiary

level qualifications. The exception is males in the Netherlands where the level of tertiary education attainment has remained stable, although it was already relatively high compared with the United Kingdom for older cohorts, being roughly 30 percent for all ages between 25 and 54. The UK has seen a noticeable increase in educational attainment in the most recent cohort to pass through tertiary education, the 20-24 year old age group. In this age group the UK has a higher percentage of tertiary educated men and women than the Netherlands. This can be explained by the fact that in the Netherlands many students are over 24 when they graduate from a university.

An interesting observation on tertiary education attainment is the relatively high level of educational attainment of elderly Dutch men. At every age group over age 60 more than one in five Dutch men has a tertiary qualification.

6.6 Population by economic activity and employment status

Figures 11 and 12 show that overall a greater percentage of the UK population are economically active compared to the population of the Netherlands. Figure 11 shows that the UK and Netherlands have similar numbers of economically active men except for below 30 years and over 60 years old, when the UK has a greater percentage classified as economically active. However, as a percentage of the whole population more men are employed in the Netherlands in every age group from age 20 up to age 50. The difference is around 3 percent, except for 20-24 years where it is 8 percent. Above 55 years the UK has higher proportions of employed men than the Netherlands.

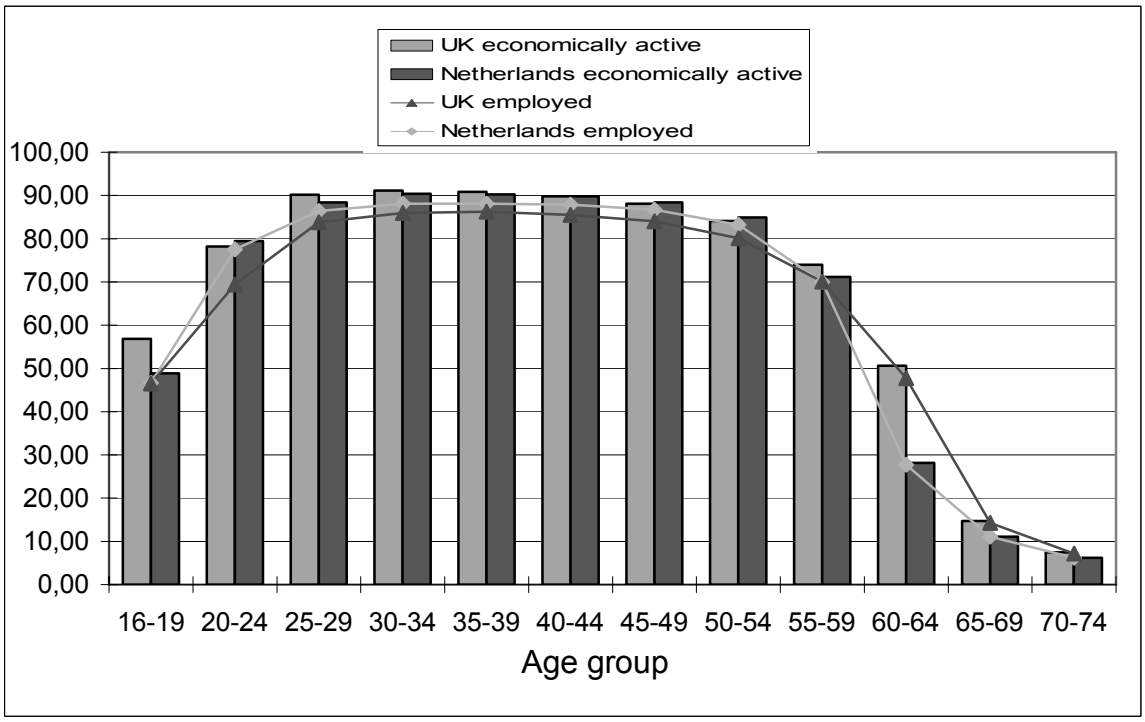

Figure 11. Economic activity and employment of the male population in the United Kingdom and the Netherlands by age group (in %)

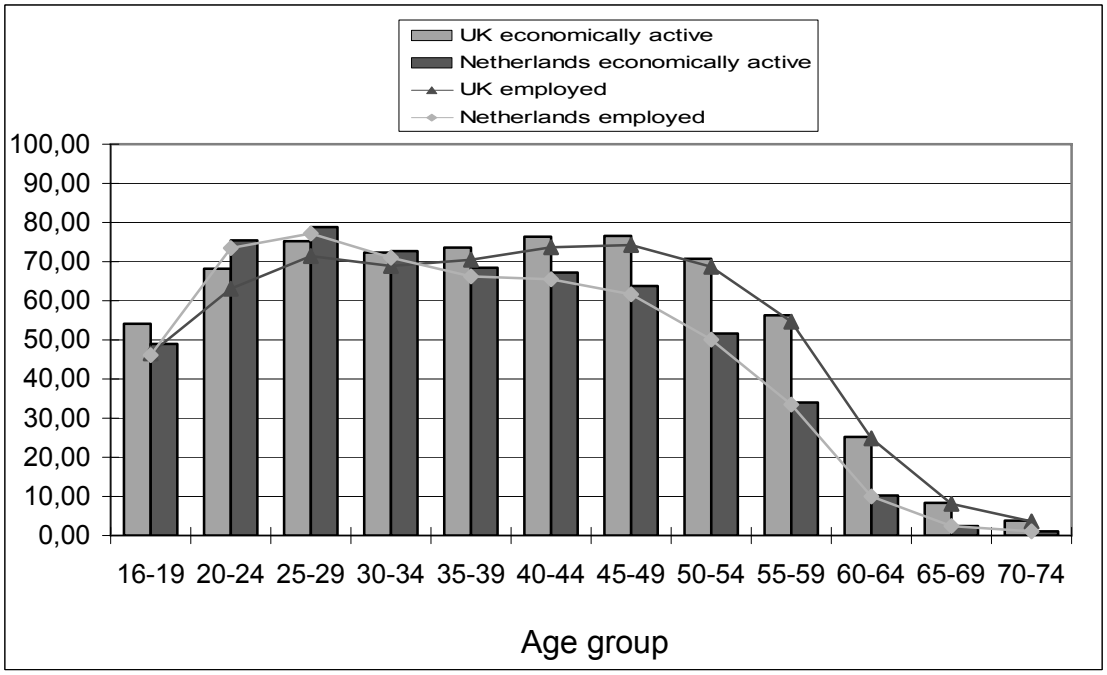

Figure 12. Economic activity and employment of the female population in the United Kingdom and the Netherlands by age group (in %)

Figure 12 shows that a greater percentage of 16-19 year old women are economically active in the UK compared with the Netherlands. The Netherlands has relatively more women aged 20-34 who are economically active. Then after age 35 at every age group there are relatively more economically active women in the UK than in the Netherlands.

The female populations of the UK and the Netherlands show quite large differences in percentages employed. Under age 30 the Netherlands has a higher percentage of employed women. From age 16-19 to age 25-29 both the UK and the Netherlands have an increasing percentage of women who are employed, the percentage of women employed then decreases at age 30-34. After age 30-34 the percentage of women employed decreases in each older age group in the Netherlands. However, in the UK the percentage of women employed increases from age 30-34 to age 45-49. Then employment levels start to decrease again but remain higher than the percentages in the equivalent age groups in the Netherlands.

So at every age after 34 years a greater percentage of women are employed in the UK than in the Netherlands, and quite a large gap opens up between female employment levels in the UK and the Netherlands. For example, at 50-54 years old only 50 percent of Dutch women are employed compared with 68 percent of UK women.

The economically active dependency ratio (for the under 16, over 60 and 16-59 economically inactive dependants) for the UK and the Netherlands are 0.90 and 0.85 respectively. This means one economically active person is supporting 0.90 economically inactive people in the UK and 0.85 economically inactive people in the Netherlands. These dependency ratios demonstrate that although Britain may have a higher proportion of economically active among those of working age than the Netherlands, these people have more economically inactive dependants than their equivalents in the Netherlands.

7. LESSONS LEARNT FROM THE 2001 CENSUS

The virtual census has proved to be a successful concept in the Netherlands. It has many advantages compared to traditional censuses. The costs are now considerably lower. Nevertheless, data on the Netherlands have become available that could be compared to results of earlier Dutch censuses and to the results of other countries that took part in the 2000 Census Round. It was the third time that the Netherlands conducted a virtual census. However, the Dutch data that have been compiled for 1981 and 1991 were of a much more limited character than the set of tables of the 2001 census. Moreover, they were largely based on a register count of the population in combination with the then existing surveys on the labour force and housing conditions.

The technique of repeated weighting has been used successfully to produce a consistent set of tables for the 2001 census. Before compiling tables with this new technique, micro-integration of the different sources in the SSD remains important. In the micro-integration process, the data are checked and incorrect data are adapted. It is strongly believed that micro-integrated data will provide more reliable results, because they are based on a maximum amount of information. Also the coverage of subpopulations will be better, because when data are missing in one source, another source can be used. Another advantage of micro-integration and repeated weighting is that there is no reason for confusion among users of statistical information: there will be one figure on each socio-economic phenomenon, instead of several figures depending on which sources have been used.

It is possible to use the technique of repeated weighting in other countries as well. However, first it should be possible to use registers for statistical purposes. In most countries, not all census variables can be derived from register information. Additional surveying then remains a necessity, but a consistent set of census tables can be produced using the technique of repeated weighting.

8. PUBLICITY ABOUT CENSUSES IN THE NETHERLANDS

At the end of 2003 the complete set of forty census tables for the Netherlands was sent to Eurostat. The book 'The Dutch Virtual Census of 2001, Analysis and Methodology' was written afterwards (Schulte Nordholt et al., 2004). This book provides a wide-ranging description of the socio-demographic and socio-economic state of the Netherlands based on the 2001 census results. It discusses differences in size and composition among households, economic activity of households, individual activity status by region, age, education level and branch of economic activity. There are separate chapters on the economic activities of young people and people of retirement age. The economic activities, levels of education and occupation of foreigners from various countries of origin are compared with each other and with the native Dutch population. Regional aspects are also examined, including commuting. The results of the 2001 census are compared with the census results of some other European countries and with earlier Dutch censuses. Lastly, the virtual census methodology used is described in some detail.

The PDF version of the book can be found at the Statistics Netherlands website, at page http://www.cbs.nl/en-GB/menu/themas/dossiers/volkstellingen/publicaties/default.htm. An extra Chapter (number 15) is available at page http://www.cbs.nl/en-GB/menu/themas/dossiers/volkstellingen/methoden/default.htm with an overview of the used data sources, methods and definitions. Hard copies of the book were sent to all authors of the book, to the management of Statistics Netherlands and to several libraries. The book was also offered to the Prime Minister, the Minister of Economic Affairs and the Minister of Education, Cultural Affairs and Science of the Netherlands and to Director-Generals of statistical offices in several countries. In August 2004, the book was publicly released at an official presentation in the Statistics Netherlands' office in Voorburg. The research process and the main findings were then presented to an audience of academics, press representatives, government officials, as well as Statistics Netherlands' employees. Several articles were written in national and regional newspapers about the Dutch virtual census of 2001 and its results. Announcements, book reviews of Schulte Nordholt et al. (2004) and interviews appeared in several journals, mailing lists and newsletters. The methodology and key results of the virtual census of 2001 were also published as Schulte Nordholt (2005).

The set of forty standard tables for the Netherlands (in Excel format) can be found at page http://www.cbs.nl/en-GB/menu/themas/dossiers/volkstellingen/publicaties/artikelen/archief/2005/default.htm

9. AVAILABILITY OF DUTCH CENSUS MICRODATA

9.1 Introduction

Protected 1 percent samples of the microdata of the Dutch censuses of 1960, 1971 and 2001 were in 2005 disseminated via the IPUMS (Integrated Public Use Microdata Series) project, see http://www.ipums.org/international. These micro datasets contain a number of demographic and economic variables and can also be analysed via the institute DANS (Data Archiving and Networked Services), see http://www.dans.knaw.nl/en/. Bona fide researchers who want to make more detailed studies on these three censuses can work on-site at the premises of Statistics Netherlands. More information about this last option can be obtained via Statistics Netherlands' Centre for Policy Research (http://www.cbs.nl/nl-NL/menu/informatie/beleid/centrum-voor-beleidsstatistiek/diensten/default.htm).

The Dutch censuses of 1960, 1971 and 2001 have been selected to be part of the IPUMS project. The censuses of 1960 and 1971 are traditional censuses, of which most of the micro data records have been recovered. The 2001 census is a virtual census, which means that it is composed of available register data and existing surveys. Unfortunately, this results in not having all variables available for all individual records. As a consequence we have not released the complete set of micro data, but an anonymised balanced sample of the individual personal records for which we have all demographic and economic variables. The sample fraction is a little bit over 1 percent of the total population.

The first stage in our cooperation in the IPUMS project has been the release of the 2001 census micro data. The selection of variables of the 2001 census has been leading in the selection of the variables of the censuses of 1960 and 1971. Due to differences in variable definitions, classifications and variable availability over time, differences among the three micro data sets remain. Also for 1960 and 1971 we release anonymised balanced 1 percent samples of the total population.

More information about the variables selected of the three censuses can be found in the next subsections. Links to more information about the 2001 census can be found in the previous section. For some more background and documentation of the 1960 and 1971 censuses we refer to the following web site: http://www.volkstellingen.nl/en/documentatie/.

9.2 The variable selection of the 2001 census

9.2.1 The sample

The sample is composed in three stages.

Of the persons 0-14 years a 1 percent sample, stratified to age (in years) and sex is drawn from the combined register data.

Out of the records from persons 15-74 years all complete records are selected. (Remember that we only had existing surveys and register data at our disposal). These records sum up to about 1 percent of the population in this age group. Record weights have been provided.

Of the records of persons 75 years or older all complete records are selected. Record weights have been provided.

Persons in institutional households are not included in the sample because they are not included in the surveys used for the virtual census of 2001.

9.2.2 The variables and their categories

Because of disclosure reasons we have to limit the detail. This results in the impossibility of providing regional detail. Moreover, some rare combinations of identifying variables will lead to a limited number of suppressions of variable scores. The specification of variables and detail of the 2001 census sample is described hereafter.

1. Sex

Male

Female

2. Age

People aged from 0 up to 79 years in 5-year classes, people aged 80 years and over in one class.

3. Household position

Child

Spouse

Cohabitant

Lone parent

Living alone

Other in private household

N.B. Spouse and Cohabitant are in two categories each: with and without children.

4. Household size

1 person
2 persons
3 persons
4 persons
5 persons
6 persons or more

5. Place of residence one year prior to the census

The Netherlands, same NUTS 3

The Netherlands, another NUTS 3 and other countries

N.a. (i.e. residents younger than 1 year)

6. Country of citizenship

The Netherlands

Rest of Europe

Rest of the world

7. Country of birth

The Netherlands

Rest of Europe

Rest of the world

8. Level of educational attainment (ISCED level 4)

Pre-primary

Primary

Lower secondary

Upper secondary

Post secondary

Tertiary

No education at all

9. Economic status

Employee, of which

Attendant at educational institutions with a job

Other employee

Employer and other employed (including self employed)

Unemployed

Attendant at educational institutions (without a job)

Retired

Engaged in family duties

Other economically inactive

10. Occupation (ISCO-COM 1 digit)

Legislators, senior officials and managers (ISCO-COM 1)

Professionals (ISCO-COM 2)

Technicians and associate professionals (ISCO-COM 3)

Clerks (ISCO-COM 4)

Service workers and shop and market sales workers (ISCO-COM 5)

Other occupations

Craft and related trades workers (ISCO-COM 7)

Plant and machine operators and assemblers (ISCO-COM 8)

Elementary occupations (ISCO-COM 9)

Not working

N.B. ISCO-COM 0 (Armed forces) and ISCO-COM 6 (Skilled agricultural and fishery workers) have been taken together (as Other occupations).

11. Branch of current economic activity (NACE, 1 letter)

A + B: Agriculture, hunting and forestry + Fishing (NACE 01-02 + NACE 05)

C + D + E: Mining and quarrying + Manufacturing + Electricity, gas and water supply (NACE 10-14 + NACE 15-37 + NACE 40-41)

F: Construction (NACE 45)

G: Wholesale and retail trade; repair of motor vehicles, motor cycles and personal and household goods (NACE 50-52)

H: Hotels and restaurants (NACE 55)

I: Transport, storage and communication (NACE 60-64)

J: Financial intermediation (NACE 65-67)

K: Real estate, renting and business activities (NACE 70-74)

L: Public administration and defence; compulsory social security (NACE 75)

M: Education (NACE 80)

N: Health and social work (NACE 85)

O: Other community, social and personal service activities (NACE 90-93)

N.B. The numbers of persons working in letters P (Private households with employed persons (NACE 95)) and Q (Extra-territorial organizations and bodies (NACE 99)) are very limited and therefore in the Dutch census of 2001 they have been divided proportionally over the other letters (branches).

12. Marital status

Single

Married

Widowed

Divorced

9.3 The variable selection of the 1971 census

9.3.1 The sample

For the census year 1971 the gross sample of 1.25 % of the total population is randomly drawn, stratified to sex, 17 age groups (16 5-year groups and 80+) and 12 regions (11 provinces and 1 region consisting of newly made land (polders) and the centrally registered population). After removing incomplete and other problematic records a net sample of over 1 percent remained. The records have been weighted to the published census combined totals of sex times age in years, 11 provinces (the newly made land and the centrally registered travelling population are added to one of the provinces) times sex times age in 5-year classes and simple totals of most of the published variables.

The total population includes persons in institutional households, this in contrast to the IPUMS dataset of the virtual census of 2001.

9.3.2 The variables and their categories

Because of disclosure reasons we have to limit the detail. This results in the impossibility of providing regional detail. Moreover, some rare combinations of identifying variables will lead to a limited number of suppressions of variable scores. The specification of variables and detail of the 1971 census sample is described hereafter.

1. Sex

Male

Female

2. Age

People aged from 0 up to 79 years in 5-year classes, people aged 80 years and over in one class.

3. Country of citizenship

The Netherlands

Other countries

The sample contains 104 records with two countries of citizenship, The Netherlands and another (European) country. These records are classified in the category The Netherlands.

Because of disclosure reasons the division of other countries into Rest of Europe and Rest of the world, is not possible, this in contrast to the 2001 census.

4. Marital status

Single

Married

Divorced or Widowed

As the category Divorced is too small in numbers in 1971, we classified these records together with the Widowed into the category Divorced or Widowed. In the 2001 census these groups can be published in two separate categories.

5. Household position

Child

Married without children

Married with children

Living alone

Other in private household

Institutional household

The category Child contains married as well as single children.

The category Other in private household includes living in servants.

The category Cohabitant (with or without children) does not exist in the 1971 census.

The category Institutional household includes living in staff.

6. Religious denomination

No religious denomination

Roman Catholic

Dutch Reformed

Reformed Churches in the Netherlands

Other Reformed

Other religious denominations

The original over 50 denominations have been combined into the above 6 categories according to their religious characteristics.

Other religious denominations include all non-Christian and all other Christian denominations.

7. *Country of birth*

The Netherlands

Rest of Europe

Rest of the world

Turkey is classified in the category Rest of Europe, as in the 2001 census.

8. Household size

1 person
2 persons
3 persons
4 persons
5 persons
6 persons or more
Institutional household

9. Economic status

Employee

Self employed

Attendant at educational institutions

Retired

Engaged in family duties

Other economically inactive

The variable economic status is made consistent with the IPUMS-dataset of the 2001 census. The original 1971 census variable Working has 16 categories, which have been reduced to the census 2001 variable categories. Records have been recoded according to the following criteria.

There is, as in the 2001 census, no lower limit in weekly working hours for persons to be classified as economically active (employee or self employed).

The distinction between employee and self employed has been made with the 1971 census variable Position in the enterprise in the main occupation. Spouses working with their husbands are classified as self employed, children working with their parents as employees.

The category Unemployed contains not enough records to be incorporated: the unemployed are classified as other economically inactive.

All persons from 4-14 years are Attendant at educational institutions. Persons aged 30 years and over are not classified as Attendant at educational institutions; the few records of persons Attendant at educational institutions with a job are classified as Attendant at educational institutions if they are under 30 years of age and as Employee if they are 30 years and over; those without a job are classified as Other economically inactive.

Retired includes only persons of age 55 and over. Retired persons without a job under 55 years of age are classified as Other economically inactive.

Engaged in family duties: persons from 4-14 are classified as Attendant at educational institutions. Persons under 4 years of age are classified as Other economically inactive. Persons of 65 years and over are classified as retired.

Other economically inactive includes unemployed persons. It does not include persons from 4-14, who are classified as Attendant at educational institutions. All persons under 4 years of age are classified as Other economically inactive.

10. Level of educational attainment

Primary school and less

Lower level

Upper lower level

Intermediate level

Tertiary level

No education at all

The 1971 census has no ISCED level variable. The harmonisation possibilities are further limited by changes in the educational system. We recoded records by means of the 1971 census variables Age, Diploma level and Educational level of persons in full time education. We point out that the level unknown suffers from selectivity, as is amplified in the aforementioned census documentation.

11. Occupation (ISCO 1 digit)

1. Professional, technical and related workers and artists (ISCO 0/1)

3. Clerical and Related Workers (ISCO 3)

4. Sales workers (ISCO 4)

5. Service workers (ISCO 5)

6. Agricultural, Animal husbandry, and Forestry Workers, Fishermen and Hunters (ISCO 6)

7. Production, and Related Workers, Transport Equipment Operators and Labourers (ISCO 7/8/9)

99. Other occupations, including occupation unknown

999. Not working

The administrative and managerial workers (ISCO 2) and the military are for disclosure reasons classified as Other occupations, including unknown.

Persons who are not working according to the variable economic status are classified as not working (code 999).

12. Branch of current economic activity (SBI 1970)

0. Agriculture + Fishing

2. Manufacturing, including mining and quarrying and electricity, gas and water supply

5. Construction and installation and fitting

6. Wholesale and retail trade; repair of motor vehicles, motor cycles and personal and household goods; hotels and restaurants

7. Transport, storage and communication

8. Financial intermediation, provision of business services

9. Other services

200 Not working

The classification of branches of economic activity is made according to the CBS 1970 Standaard bedrijfsindeling (SBI), which corresponds with the Nomenclature statistique des activités économiques dans la Communauté Européenne (NACE 1970) and the United Nations International Standard Industrial Classification of all Economic Activities (ISIC 1968).

Mining and quarrying and electricity, gas and water supply have been combined with Manufacturing because of their small numbers.

Persons who are according to the variable economic status retired or attendant at educational institutions are classified as not working.

9.4 The variable selection of the 1960 census

9.4.1 The sample

For the census year 1960 the gross sample of 1.25 percent of the total population has been drawn randomly, stratified to sex, 17 age groups (16 5-year groups and 80+) and 12 regions (11 provinces and 1 region consisting of newly made land (polders) and the centrally registered population). After removing incomplete and other problematic records a net sample of over 1 percent remained. The records have been weighted to the published census combined totals of sex times age in years and region times sex times age in 5-year classes.

The total population includes persons in institutional households, this in contrast to the virtual census of 2001.

The source material is the over 11 million original punch cards, which have been reread and digitised from 1973 onwards. A report in English on the reconstruction of the dataset is available.

9.4.2 The variables and their categories

Because of disclosure reasons we have to limit the detail. This results in the impossibility of providing regional detail. Moreover, some rare combinations of identifying variables will lead to a limited number of suppressions of variable scores. The specification of variables and detail of the 1960 census sample is described hereafter.

1. Sex

Male

Female

2. Age

People aged from 0 up to 74 years in 5-year classes, people aged 75 years and over in one class.

3. Marital status

Single

Married

Divorced or Widowed

The Dutch law has the facility to be married and officially "separated from table and bed": persons in this category in the 1960 census are attributed to the IPUMS category Divorced. As the category Divorced, including separated from table and bed is too small in numbers in 1960, we classified these records together with the Widowed into the category Divorced or Widowed. The same procedure is followed in 1971. In the 2001 census these groups can be published separately. The few records with Marital status unknown are classified as Single.

4. Household position

Child

Married without children

Married with children

Living alone

Other in private household

Institutional household

The category Child contains only single children. Other children are classified as Other in private household.

Other in private household includes living in servants.

The category Cohabitant (with or without children) does not exist in the 1960 census.

The category Institutional household includes staff living in.

5. Religious denomination

No religious denomination

Roman Catholic

Dutch Reformed

Reformed Churches in the Netherlands

Other Reformed

Other religious denominations

The original over 30 denominations have been combined into the above 6 categories according to their religious characteristics. Other religious denominations include all non-Christian and all other Christian denominations.

6. Country of birth

The Netherlands

Rest of Europe

Rest of the world

The above three categories are combinations (because of disclosure reasons) of the original 13 categories.

7. Economic status

Employee

Self employed

Attendant at educational institutions

Retired

Engaged in family duties

Other economically inactive

The variable economic status is made consistent with the IPUMS dataset of the 2001 census. The categories of the original 1960 census variable Professional position have been reduced to the census 2001 categories. Records have been recoded according to the following criteria.

Spouses working with their husbands are classified as self employed, children working with their parents as employees.

The category Unemployed contains not enough records to be incorporated: the unemployed are classified as Other economically inactive.

All persons from 4-14 years are Attendant at educational institutions. Persons aged 30 years and over are not classified as Attendant at educational institutions; the few records of persons Attendant at educational institutions with a job are classified as Attendant at educational institutions if they are under 30 years of age and as employee if they are 30 years and over; those without a job are classified as Other economically inactive.

Retired includes only persons of age 55 and over. Retired persons without a job under 55 years of age are classified as Other economically inactive.

Engaged in family duties: persons from 4-14 are classified as Attendant at educational institutions. Persons under 4 years of age are classified as Other economically inactive. Persons of 65 years and over are classified as Retired.

Other economically inactive includes unemployed persons. It does not include persons from 4-14, who are classified as Attendant at educational institutions. All persons under 4 years of age are classified as Other economically inactive.

8. Level of educational attainment

Lower level and less, including unknown level

Upper lower level

Intermediate level and higher

The 1960 census has no ISCED-level variable (the ISCED did not yet exist in 1960). The harmonisation possibilities are further limited by changes in the educational system. We recoded records by means of the 1960 census variables Age, General education and Professional education.

9. Occupation

Professional, technical and related workers and artists

Clerical and Related Workers

Sales workers

Agricultural, Animal husbandry, and Forestry Workers, Fishermen and Hunters

Transport Equipment Operators

Production, and Related Workers, and Labourers

Service workers, sports and recreational workers

Other occupations, including occupation unknown

Not working

The CBS 1960 census occupation classification conforms to the ISCO. The administrative and managerial workers and the military are for disclosure reasons classified as Other occupations, including unknown.

10. Branch of current economic activity (SITC)

Agriculture + Fishing

Manufacturing, including mining and quarrying and electricity, gas and water supply

Construction and installation and fitting

Wholesale and retail trade; financial intermediation

Transport, storage and communication

Other services

Not working

The CBS 1960 census classification of branches of economic activity is made according to the United Nations Ecosoc International Standard Industrial Classification of all Economic Activities (ISIC).

The categories Mining and quarrying and Electricity, gas and water supply have been combined with the category Manufacturing because of their small numbers. The same procedure was executed for the 1971 IPUMS dataset.

Persons who are according to the variable economic status retired or attendant at educational institutions are classified as not working.

Household size

The variable household size is in the 1960 census known for heads of households only and can thus not be provided in the IPUMS dataset.

Country of citizenship

The variable Country of citizenship cannot be supplied because of the very small numbers of people with foreign citizenship.

10. CONSIDERATIONS FOR FURTHER HARMONISATION OF SOME VARIABLES AND RECOMMENDATIONS FOR FUTURE CENSUS ROUNDS

In this section, based on the experiences of the Dutch virtual census of 2001, some recommendations are made for future Census Rounds. It is very useful to have census guidelines and a table program, but some errors in this program have to be corrected and some aspects have to be described in some more detail for the next Census Round. The guidelines have originally been written with the idea in mind that all countries conduct a traditional census, but more and more countries have chosen other options. In revising the guidelines for the 2010 Census Round, this aspect has been taken into account to a certain extent. It is a challenge to change the table program according to the remarks made below.

The number of different classifications for one variable (e.g. age) is sometimes too large in the table program. This implies that sometimes cells in tables cannot be estimated consistently, whereas less estimation problems would be occurred if the number of different classifications would have been reduced. This problem is much more severe if non-nested classifications of the same variable are used.

What to do with people who have different economic activities at the same time? In a traditional census one can ask the respondent about his or her main economic activity. In register-based censuses one has to find a criterion to choose one economic activity per person. If different countries use different criteria, the comparability of the results will be hampered. In the Netherlands we could not distinguish between employers and other employed people and therefore these two self-employed categories have been merged. People who were at the same time employed and self-employed were counted as employees. As we have no register of unemployed people, this group is the only part of the economically active population that has to be estimated. Therefore, sometimes the number of economically active people is an estimate, although we know exactly how many people are employee or self-employed. People who are at the same time both economically active and student are counted as economically active people. The population of retired people is not clearly defined. What to do with people who are partly retired? And what with people who live from their property instead of from their pension? It is necessary to make international priority rules on the variable economic activity to improve the comparability of the results in different countries.

More attention has to be paid to the variable country of birth and less to the variable citizenship in the next Census Round. Different countries have namely different policies towards changing nationality. International more relevant comparisons can be made by distinguishing first and second generation foreigners on the basis of the variables country of birth of the person and his or her parents. This is done in the PAU (Population Activities Unit of the UN) tables and Chapter six of Schulte Nordholt et al. (2004). In the family nuclei it is

useful to add extra categories for same sex (married or cohabiting) couples. This is a growing group that has to be taken into account in the next Census Rounds. For the variable family status the rest category 'child with other family status' is missing.

The NUTS classifications change over time. Therefore, it is crucial to provide the countries with the table lay-outs including the classifications of the census year. The list of country codes has to be improved and used consistently in all tables. Now only a general unknown category is included. This implies that for countries not in the list all the records are counted as unknowns, which implies that the totals per continent become incorrect. What to do with no longer existing countries not in the list? This is especially relevant for the variable country of birth if the original country was split.

For the NACE codes it is useful to add a separate category unknown. Now we have attributed the unknowns proportionally to the other categories. Also for the variable occupation a category unknown should be introduced. For 2001 we have included the unknowns in the total, but that implies that the sum of the occupation categories is often smaller than the total. Finally, for the ISCED a category unknown is included in the table, but here we merged the unknowns with the categories no education at all and level of education pre-primary as we could not distinguish among these three categories.

When we take the lessons learnt on the importance of nested (age) classifications, the priority rules for economic activities, the variable country of birth, and on the NUTS and other classifications into account for the table program, the comparability of the results of the different countries will be improved enormously.

11. APPLICATIONS OF STATISTICAL DISCLOSURE CONTROL METHODS

11.1 Introduction

National Statistical Institutes (NSIs) and Market Research Bureaus conduct surveys about many different topics. To reach this aim they have developed a fully-equipped statistical production process. It is a long way from collecting raw data to publishing public information.

The information from statistics becomes available for the public in tabular and microdata form. Historically, only tabular data were available and NSIs had a monopoly on the microdata. Since the eighties the PC revolution leaded to the end of this monopoly. Now also other users of statistics have the possibility of using microdata. These microdata can be conveyed with floppies, CD-ROMs and other means. Recently also other possibilities of getting statistical information have become more popular as remote access and remote execution. With these techniques researchers can get access to data that remain in a statistical office or can execute set-ups without having the data on their own PC. For very sensitive information some NSIs have the possibility to let bona fide researchers work on-site within the premises of the NSI.

The task of statistical offices is to produce and publish statistical information about society. The data collected are ultimately released in a suitable form to policy makers, researchers and the general public for statistical purposes. The release of such information may have the undesirable effect that information on individual entities instead of on sufficiently large groups of individuals is disclosed. The question then arises how the information available can be modified in such a way that the data released can be considered statistically useful and do not jeopardize the privacy of the entities concerned. The Statistical Disclosure Control theory is used to solve the problem of how to publish and release as much detail in these data as possible without disclosing individual information (Willenborg and De Waal, 1996 and 2001).

This section is partly based on earlier work on Statistical Disclosure Control (e.g. Schulte Nordholt, 2001). It discusses the available methods to protect sensitive information. The tables produced by statistical offices on the basis of the microdata of surveys have to be protected against the risk of disclosure. Therefore the software package τ-ARGUS (Hundepool et al, 2003a) can be applied on the tables produced. More information about τ-ARGUS and how this package can be applied are given in subsection 11.2. Subsection 11.3 explains how microdata for research and public use microdata files can be produced using the software package μ-ARGUS (Hundepool et al, 2003b). The option for bona fide researchers to work on-site at Statistics Netherlands on richer microdata files is explained in subsection

11.4. Also other methods that allow use of data are discussed in that chapter. Finally, in subsection 11.5 a discussion follows about the current state and some possible extensions for the ARGUS packages. Some conclusions are also drawn in subsection 11.5. Many of the ideas for this paper came from Citteur and Willenborg (1993), Groot and Citteur (1997), Willenborg (1993) and Willenborg and De Waal (1996 and 2001).

The software packages τ-ARGUS and μ-ARGUS have emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth Framework Programme of the European Union. The Computational Aspects of Statistical Confidentiality (CASC) project can be seen as a follow-up of the SDC project. The CASC project is funded under the Fifth Framework Programme for Research, Technological Development and Demonstration (RTD) of the European Union. It builds further on the achievements of the SDC project. On the other hand it has new objectives. It concentrates more on practical tools and research needed to develop them. In the CASC project fourteen partners from five different European countries (Germany, Italy, the Netherlands, Spain and the United Kingdom) work closely together. One of the main tasks of this consortium is to further develop the ARGUS-software which has been put in the public domain by the SDC project consortium. The CASC project involves both research and software development. As far as research is concerned the project concentrates on those areas that can be expected to result in practical solutions, which are then being built into the software. The CASC project has been designed around the software twin ARGUS. This will make the outcome of the research readily available for application in the daily practice of National Statistical Institutes and Market Research Bureaus. More information about the CASC project can be found in Hundepool (2001).

11.2 The release of tabular data

Many tables are produced on the basis of surveys. As these tables have to be protected against the risk of disclosure, the software package τ-ARGUS (Hundepool et al, 2003a) can be applied. Two common strategies to protect against the risk of disclosure are table redesign and the suppression of individual values. It is necessary to suppress cell values in the tables because publication of (good approximations of) these values may lead to disclosure. These suppressions are called primary suppressions.

A dominance rule is often used to decide which cells have to be suppressed. This rule states that a cell is unsafe for publication if the n major contributors to that cell are responsible for at least p percent of the total cell value. The idea behind this rule is that in unsafe cells the major contributors can determine with great precision the contribution of their competitors. In τ-ARGUS the default value for n is 3 and the default value for p is 70 %, but these values can be changed easily if the user of the package prefers other values. Using the chosen dominance

rule τ-ARGUS shows the user which cells are unsafe. In publications crosses (×) normally replace unsafe cell values. Other rules that can be used to decide which cells have to be suppressed are the p-percent rule and the pq rule. The p-percent rule states that approximate disclosure of magnitude data (business data reporting non-negative quantities about certain establishments or similar entities) occurs if the user can estimate the reported value of some respondent too accurately. Such disclosure occurs, and the table cell is thus declared sensitive, if upper and lower estimates for the respondent's value are closer to the reported value than a prespecified percentage, p. In the derivation for the p-percent rule, one assumes that there was a limited prior knowledge about respondent's values. Some people believe that agencies should not make this assumption. In the pq rule, agencies can specify how much prior knowledge there is by assigning a value q, which represents how accurately respondents can estimate another respondent's value before any data are published ($p<q<100$).

The most widespread technique used to identify sensitive cells is the dominance rule. The p-percent rule can be considered as a special kind of pq rule. The pq rule is intuitively clearer and easier to extend in specific situations than the dominance rule. The pq rule can also be used if we have negative contributions or cell values in the table. When some of the contributors know approximately some of the other contributions to a cell value, this prior information can be taken into account with the pq rule. This is not the case with the dominance rule. An example of such a situation is when permission is obtained from a respondent in a sensitive cell to publish the cell. Such a waiver can be useful for publication purposes and not too demanding for a large public company where similar information is already in the public domain. The pq rule can handle waivers whereas with the dominance rule it is not clear how to continue as it should not be allowed to disclose approximately the value of another contributor to that cell. Finally, the pq rule has the advantage that both upper and lower limits are taken into account whereas when the dominance rule is used, only an upper limit can be deducted. The last mentioned disadvantage for the dominance rule also holds for the p-percent rule. In spite of these disadvantages not many countries have already experience in using other rules than the dominance rule for the identification of sensitive cells in tables. When the p-percent rule and the pq rule will become available in standard software packages for Statistical Disclosure Control it can thus be expected that these rules will become more popular.

As marginal totals are given as well as cell values, it is necessary to suppress further cells in order to ensure that the original suppressed cell values cannot be recalculated from the marginal totals. Even if it is not possible to recalculate the suppressed cell value exactly, it is often possible to calculate it within a sufficiently small interval. In practical situations every cell value is often non-negative and thus cannot exceed the marginal totals in the row or column. If the size of such an interval is small, then the suppressed cell can be estimated with great precision, which is of course undesirable. Therefore, it is necessary to suppress additional cells to ensure that the intervals are sufficiently large. A user has to indicate how large a sufficiently large interval should be. This interval is called the safety range and a

safety range could e.g. have a lower bound of 70 % and an upper bound of 130 % of the cell value. A user of a table cannot see if a suppression is a primary or secondary suppression: normally all suppressed cells are indicated by crosses (×). Not revealing why a cell has been suppressed helps to prevent the disclosure of information.

Preferably the secondary suppressions are executed in an optimal way, however the definition of optimal is an interesting problem. Several measures for the loss of information can be defined and then the loss of information according to the measure chosen should be minimised. Four possibilities are:

- the minimisation of the number of secondary suppressions;
- the minimisation of the total of the suppressed values;
- the minimisation of the total number of individual contributions to the suppressed cells;
- the minimisation of a weighted function of scores attributed to cells that symbolise information, where empty cells get weight 0 and neighbouring cells to primary suppressions get lower weights than cells further away from primary suppressions.

Often, the minimisation of the number of secondary suppressions is considered to be optimal. Also the possibilities to minimise the total of the suppressed values or the total number of individual contributions to the suppressed cells are now and then used. The minimisation of the total of the suppressed values is of course only relevant if all cell values are non-negative. For the last mentioned minimisation one can take the hierarchy of the table into account and then software tailored to the specific needs is required. In τ-ARGUS the option of minimising the total of the suppressed values has been implemented as the default. In τ-ARGUS version 2.1 it is also possible to minimise the total number of individual contributions to the suppressed cells. If that criterion is used a so-called cost variable that is equal to 1 for every record is used to execute the secondary suppressions. Also the option of minimising the number of secondary suppressions itself has been implemented in τ-ARGUS version 2.1. This implies that with τ-ARGUS version 2.1 the three options implemented that may lead to different resulting groups of secondary suppressions can be compared.

If the process of secondary suppressions is directly executed on the most detailed tables available, large numbers of local suppressions will often result. Therefore, it is better to try to combine categories of the spanning (explanatory) variables. A table redesigned by collapsing strata will have a diminished number of rows or columns. If two safe cells are combined a safe cell will result. If two cells are combined when at least one is not safe it is impossible to say beforehand if the resulting cell will be safe or unsafe, but this can easily be checked afterwards by τ-ARGUS. However, the remaining cells with larger numbers of enterprises tend to protect the individual information better, which implies that the percentage of unsafe cells tends to diminish by collapsing strata. Thus, a practical strategy for the protection of a

table is to start by combining rows or columns. This can be executed easily within τ-ARGUS. Small changes in the spanning variables can most easily be executed by manual editing in the recode box of τ-ARGUS, while large changes can be handled more efficiently in an externally produced recode file which can be imported into τ-ARGUS without any problem. After the completion of this redesign process, the local suppressions can be executed with τ-ARGUS given the parameters for *n*, *p* and the lower and upper bound of the safety range.

As normally many tables are produced on the basis of a survey and the software package used for the data protection is based on individual tables, there is the risk that although each table is safe, the combination of the data in these tables will disclose individual information. This may be the case when the tables have spanning and response variables in common. Version 2.1 of τ-ARGUS does support linked tables. An earlier version had an option to protect such tables, but this was not warranted. This implies that the aim is now reached to have extended τ-ARGUS in such a way that it is able to deal with an important sub-class of linked tables, namely hierarchical tables. A hierarchical table is an ordinary table with marginals, but also with additional subtotals. Hierarchical tables imply much more complex optimisation problems to be solved than single tables. Some approximation methods exist for finding optimal solutions for these problems. The extension version 2.1 of τ-ARGUS was released in the CASC (Computational Aspects of Statistical Confidentiality) project.

11.3 The release of microdata for researchers and public use microdata files

Many users of surveys are satisfied with the safe tables released by statistical offices. However, some users require more information. For many surveys microdata for researchers are released. The software package μ-ARGUS (Hundepool et al, 2003b) is of help in producing these microdata for researchers. For the microdata for researchers Statistics Netherlands uses the following set of rules:

1. Direct identifiers should not be released.
2. The indirect identifiers are subdivided into extremely identifying variables, very identifying variables and identifying variables. Only direct regional variables are considered to be extremely identifying. Each combination of values of an extremely identifying variable, a very identifying variable and an identifying variable should occur at least 100 times in the population.
3. The maximum level of detail for occupation, firm and level of education is determined by the most detailed direct regional variable. This rule does not replace rule 2, but is instead an extension of that rule.

4. A region that can be distinguished in the microdata should contain at least 10 000 inhabitants.
5. If the microdata concern panel data direct regional data should not be released. This rule prevents the disclosure of individual information by using the panel character of the microdata.

In the case of most Statistics Netherlands' business statistics the responding enterprises are obliged by a law on official statistics to provide their data to Statistics Netherlands. This law dates back to 1936 and was renewed in 1996 without changing the obligation of enterprises to respond. No individual information may be disclosed when the results of these business surveys are published. The law states that no microdata for research may be released from these surveys. Statistics Netherlands can therefore provide two kinds of information from these surveys: tables and public use microdata files. Public use microdata files contain much less detailed information than microdata for research. The software package μ-ARGUS (Hundepool et al, 2003b) is also of help in producing public use microdata files. For the public use microdata files Statistics Netherlands uses the following set of rules:

1. The microdata must be at least one year old before they may be released.
2. Direct identifiers should not be released. Also direct regional variables, nationality, country of birth and ethnicity should not be released.
3. Only one kind of indirect regional variables (e.g. the size class of the place of residence) may be released. The combinations of values of the indirect regional variables should be sufficiently scattered, i.e. each area that can be distinguished should contain at least 200 000 persons in the target population and, moreover, should consist of municipalities from at least six of the twelve provinces in the Netherlands. The number of inhabitants of a municipality in an area that can be distinguished should be less than 50 % of the total number of inhabitants in that area.
4. The number of identifying variables in the microdata is at most 15.
5. Sensitive variables should not be released.
6. It should be impossible to derive additional identifying information from the sampling weights.
7. At least 200 000 persons in the population should score on each value of an identifying variable.
8. At least 1 000 persons in the population should score on each value of the crossing of two identifying variables.

9. For each household from which more than one person participated in the survey we demand that the total number of households that correspond to any particular combination of values of household variables is at least five in the microdata.
10. The records of the microdata should be released in random order.

According to this set of rules the public use files are protected much more severely than the microdata for research. Note that for the microdata for research it is necessary to check certain trivariate combinations of values of identifying variables and for the public use files it is sufficient to check bivariate combinations. However, for public use files it is not allowed to release direct regional variables. When no direct regional variable is released in a microdata set for research, then only some bivariate combinations of values of identifying variables should be checked according to the Statistical Disclosure Control rules. For the corresponding public use files all the bivariate combinations of values of identifying variables should be checked.

The software package μ-ARGUS is of help to identify and protect the unsafe combinations in the desired microdata file. Thus rule 2 for the microdata for researchers and the rules 7 and 8 for the public use microdata files can be checked with μ-ARGUS. Global recoding and local suppression are two data protection techniques used to produce safe microdata files. In the case of global recoding several categories of an identifying variable are collapsed into a single one. This technique is applied to the entire data set, not only to the unsafe part of the set, so that a uniform categorisation of each identifying variable is obtained.

In the field of microdata several new techniques are being investigated in the CASC (Computational Aspects of Statistical Confidentiality) project. New methodologies like post randomisation (PRAM), micro-aggregation and noise-addition will be implemented in new versions of μ-ARGUS that will be released in the near future. PRAM is a perturbative method for disclosure protection of categorical variables (see e.g. Gouweleeuw, Kooiman, Willenborg and De Wolf, 1998). Applying PRAM means that for each record in a microdata file the score on one or more categorical identifying variables may be misclassified into different scores according to a predetermined probability mechanism (Van den Hout and Van der Heijden, 2002). Since the original data file is perturbed, it will be difficult for an intruder to identify records with certainty as corresponding to certain individuals in the population. In other words, the randomness of the procedure implies that matching a record in the perturbed file to a record of a known individual in the population could, with a high probability, be a mismatch. The records in the original file are thus protected, which is the main goal of applying PRAM. On the other hand, since the probability mechanism that is used when applying PRAM is known, characteristics of the true data can be estimated from the perturbed data file. Hence, it is still possible to perform all kinds of statistical analyses after PRAM has been applied. However, using the transition matrix with the misclassification probabilities to take into account the perturbation due to PRAM requires extra effort and becomes more

complex when the research questions become more complex. Two key questions concerning PRAM are the following:

- How should the misclassification probabilities be chosen in order to make the released microdata file safe?
- How should statistical analyses be adjusted in order to take into account the misclassification probabilities?

The first question is addressed in Willenborg and De Waal (2001) and the second question in Van den Hout and Van der Heijden (2002).

The implementation of the new methodologies in μ-ARGUS will allow for experimenting with these techniques. In the future a mixture of several disclosure protection methods can be applied, e.g. combining PRAM, global recoding and local suppression. Generally speaking, it is at the moment still unclear what the implications will be for Statistical Disclosure Control rules in official statistics.

To measure the quality of the methods applied disclosure risk and information loss models will be implemented in new versions of μ-ARGUS too. A disclosure risk model is specified to distinguish safe from unsafe microdata. Disclosure models can differ greatly in their levels of sophistication. In a fairly simple disclosure model a combination of values is safe only if the estimated frequency of its occurrence in the population is above a certain threshold value. Which combinations to consider is also part of the disclosure risk model that one applies, and should be specified by the data protector. Fienberg and Makov (1998) and Skinner and Holmes (1998) have proposed more advanced disclosure risk models. They use log linear models for the estimation of the individual risk. Whatever disclosure risk model is used, one always has to make some assumptions on the nature of possible attacks by intruders to the privacy of an individual (see e.g. Keller and Bethlehem, 1992, Mokken et al, 1992, Elliot and Dale, 1999 and Elliot, 2001).

If unsafe microdata are going to be transferred into safe microdata it is necessary to have a measure of information loss at one's disposal. This measure is used to limit the amount of damage done to the microdata when they are being modified by the data protector. In case of applying local suppressions, μ-ARGUS uses the number of local suppressions as measure of information loss. The more suppressions the higher the information loss. The optimisation problem that has to be solved is to select the local suppressions in such a way that the resulting microdata are safe and the associated information loss is minimised. As μ-ARGUS uses the number of local suppressions as measure of information loss, the problem how to suppress in an optimal way can be solved record-wise (De Waal and Willenborg, 1998). Another possibility is to choose the number of different categories affected by the suppression as measure of information loss. The minimisation problem then tends to be more difficult to

solve in practice. In some cases this problem can be decomposed into a number of smaller, and therefore easier to solve, problems. In case of global recoding of an identifying variable the information loss depends on the valuation of the importance of the variable and the valuation of each of the possible predefined codings for the variable. Optimisation models for a special form of global recoding have been developed by Hurkens and Tiourine (1998). Both global recoding and local suppression lead to information loss, because either less detailed information is provided or some information is not given at all. A balance between global recoding and local suppression should always be found in order to make the information loss due to the Statistical Disclosure Control measures as low as possible. It is recommended to start by recoding some variables globally until the number of unsafe combinations that has to be protected is sufficiently low. Then the remaining unsafe combinations have to be protected by local suppressions.

11.4 Other methods that allow use of data

All techniques described in chapter 3 necessarily involve data manipulation or suppression and are likely to reduce the quality of estimates to be produced from the data. As a result, National Statistical Institutes (NSIs) have begun to investigate other methods that allow use of data while protecting confidentiality of sensitive information given by respondents. These methods allow the data to be used in an environment controlled by the NSI and require that its use be subject to the same legal and ethical protections placed on the NSI itself.

Some NSIs (e.g. in the U.S.A.) have introduced the process of licensing whereby institutions and researchers outside the NSIs temporarily gain access to (a part of the) data at their site by agreement to conform to legal protections surrounding those data that are imposed on the NSI. Data licensing is thus a way to provide access to data when they cannot be released to the public because of confidentiality concerns. It is necessary that periodic inspections are performed of the licensed sites. Also a good organisation of the licensed files within the NSI is a necessity for the agreement to become a success.

Probably the most important access modality developed in the past decade is that of restricted access sites. These sites permit NSIs to respond to the microdata needs of researchers. Some researchers need namely more information than is available in the released microdata for researchers or public use microdata files. As the releasing of richer data is not allowed, it is then possible for individual researchers to perform their research on richer microdata on the premises of the NSIs. Statistics Netherlands is one of the NSIs that has such a facility. Bona fide researchers have the opportunity to work on-site in a secure area within Statistics Netherlands. Researchers can choose at will between the two locations of Statistics Netherlands: Voorburg in the west of the Netherlands and Heerlen in the south of the Netherlands. The possibility to export any information is however only possible with the

permission of the responsible statistical officer. They can apply standard statistical software packages and also bring their own programmes. Like all employees of Statistics Netherlands, these people who work on-site have to swear an oath to the effect that they will not disclose the individual information of respondents (Kooiman, Nobel and Willenborg, 1999).

The researchers who work on-site on Statistics Netherlands' economic data have to take the rules of the Centre for Research of Economic Microdata (CEREM) into account. The most important rules are:

- researchers must be associated with a recognised research institute (e.g. a university);
- there must be a research proposal that conforms to current scientific standards;
- the researcher and his superior have to sign a confidentiality warrant;
- the researcher obtains only access to the data needed for his project;
- the data do not contain information on names and addresses of the enterprises;
- data related to the two most recent years will not be supplied;
- it is forbidden to let data or not safeguarded intermediate results leave the premises of Statistics Netherlands;
- all prospective publications will be screened with respect to the risk of disclosure;
- all publications will be in the public domain;
- a public register contains the researcher's name(s), the research project, the publication(s) and the databases provided.

The facility provided by Statistics Netherlands is not free of charge. As a rule the researcher has to pay the cost for the supply of the required data. In addition, there is a tariff for using the on-site facility.

Finally, an option is to allow remote access. This access modality combines the advantage of licensing that researchers can stay in their own institute and the advantage of working on site that the data stay in the NSI. Normally, researchers get access through an intermediary controlled by the NSI that guarantees that all use conforms to the law. One step further goes the option of remote execution. Then no longer an intermediary is placed between the researcher and the NSI. With remote execution researchers can execute set-ups without having the data on their own PC. Although remote execution is a more efficient option than remote access the question is whether the security systems are strong enough to let this technique become an often used modality. Currently, Statistics Netherlands has a Centre for Policy Research that is running a successful pilot with the Ministry of Social Affairs and Employment. In this pilot the remote execution is limited in the sense that employees of

Statistics Netherlands still check manually the set-ups that are sent to the Centre for Policy Research.

11.5 Discussion and conclusions

The software packages τ-ARGUS and μ-ARGUS have emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth RTD Framework Programme of the European Union. These software packages appear to be of great help in the practice of Statistical Disclosure Control. Many of the protection problems of statistical data can be solved using the ARGUS packages.

The manuals (Hundepool et al, 2003a and b) are of great help for the users of the ARGUS packages. However, there are always additional things to desire. In the case of τ-ARGUS more research is needed into how a set of linked tables and corresponding metadata could efficiently be dealt with in an automated way. More research is also needed into how consecutive years of the same survey can be protected from disclosure. In the case of μ-ARGUS, it is important to clarify in the package the difference between protecting microdata for research and protecting public use microdata files. As μ-ARGUS can be used with lots of different protection criteria, it is important to help the users to understand how different strategies can be executed using the package. Recently, research has been directed at perturbation methods by adding stochastic noise to microdata. It would be good to have several options in μ-ARGUS to perturb data as a protection technique. Especially interesting for a user is to know how well protected microdata are after the perturbation process.

It can be concluded that there is still a lot of research to be done in the field of Statistical Disclosure Control. Hopefully, new versions of the ARGUS packages (that include results of the on-going research) will soon be released to the user community. The production of these new versions is part of the CASC project (see Hundepool, 2001). To promote the results of the statistical projects under the Fourth RTD Framework Programme of the European Union the AMRADS (Accompanying Measures in Research And Development in Statistics) project is funded under the Fifth RTD Framework Programme. Many courses and conferences are being organised, among other topics, about Statistical Disclosure Control. These activities will stimulate the progress in the implementation of Statistical Disclosure Control methods and techniques in many different countries.

A couple of National Statistical Institutes (NSIs) and Market Research Bureaus have their own research activities (see e.g. Bethlehem and Pannekoek, 1998). Besides those NSIs the statistical agency of the European Union, Eurostat, has established itself as a main promoter of research in statistics. A dedicated budget for subsidising targeted research and development projects has been available in recent years. Many projects (e.g. the projects SDC, CASC and AMRADS discussed in this paper) were subsidised by the European Union. Before the Fifth RTD Framework Programme of the European Union started, the European Plan for Research in Official Statistics (EPROS) was launched. Statistical Disclosure Control is one of the topics

mentioned in EPROS. Eurostat has stimulated the forming of consortia of researchers from Universities, NSIs and Market Research Bureaus. This way, many ideas have been exchanged and many researchers learnt a lot form each other. Not all subsidised projects always lead to good results that can be implemented in practice. However, it is hard to predict which projects will become most successful. Critical success factors are at any rate a clear aim of the project and an efficient project organisation. Hopefully, Eurostat (and maybe also other international organisations) will continue to find ways to stimulate research in statistics in the future as well. Although one never knows exact outcomes of research projects beforehand, it is clear that subsidising statistical research projects has led to economies of scale and speeded up the process towards better and more comparable statistics.

In this paper methods have been described that have been developed to protect confidentiality, while at the same time providing access to data, through various means that either alter the data or restrict access to them. The balance between data confidentiality and data access is a delicate one. Hopefully, the new research methods and software for Statistical Disclosure Control can help in keeping the right balance.

As part of the CASC project new versions of the ARGUS packages become available for users. New manuals for τ-ARGUS and μ-ARGUS became available (Hundepool et al, 2003a and b) that have been tested intensively as part of the CASC project. Both manuals are of great help to the testers. In the new version of τ-ARGUS HiTaS has been included, so that from this version onwards hierarchical tables can be dealt with as well. In the new version of μ-ARGUS new options are PRAM and individual risk models. The ARGUS packages have moved towards interfaces with several state of the art engines produced by statisticians from many different countries. The most recent information is published at the CASC website: http://neon.vb.cbs.nl/casc.

REFERENCES ON CENSUSES

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

Corbey, P., 1994, "Exit the population Census", *Netherlands Official Statistics*, 9, summer 1994, pp. 41-44.

Duin, C. van and V. Snijders, 2003, "Simulation studies of repeated weighting", *Discussion paper 03008*, Statistics Netherlands, Voorburg / Heerlen. http://www.cbs.nl/en-GB/menu/methoden/research/discussionpapers/archief/2003/default.htm

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

Houbiers, M., 2004, "Towards a social statistical database and unified estimates at Statistics Netherlands", *Journal of Official Statistics*, 20, pp. 55-75.

Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen and V. Snijders, 2003, "Estimating consistent table sets: position paper on repeated weighting", *Discussion paper 03005*, Statistics Netherlands, Voorburg / Heerlen. http://www.cbs.nl/en-GB/menu/methoden/research/discussionpapers/archief/2003/default.htm

Kroese, A.H. and R. H. Renssen, 2000, "New applications of old weighting techniques, constructing a consistent set of estimates based on data from different sources", *ICES II, Proceedings of the second international conference on establishment surveys, survey methods for businesses, farms, and institutions, invited papers*, June 17-21, 2000, Buffalo, New York, American Statistical Association, Alexandria, Virginia, United States, pp. 831-840.

Laan, P. van der, 2000, "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, pp. 7-15.

OECD, 2003, "Education at a glance", *OECD Publications*, Paris, France.

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

Rao, J.N.K., 2003, *Small area estimation*, Wiley, New York, United States.

Schulte Nordholt, E., 2005, "The Dutch virtual Census 2001: A new approach by combining different sources", *Statistical Journal of the United Nations Economic Commission for Europe*, 22, 2005, pp. 25-37.

Schulte Nordholt, E., M. Hartgers and R. Gircour (Eds.), 2004, *The Dutch Virtual Census of 2001, Analysis and Methodology*, Statistics Netherlands, Voorburg / Heerlen, July, 2004. http://www.cbs.nl/NR/rdonlyres/D1716A60-0D13-4281-BED6-3607514888AD/0/b572001.pdf

Statistics Netherlands, 2003, "Urban Audit II, the implementation in the Netherlands", *Report, BPA no. 2192-03-SAV/II*, Statistics Netherlands, Voorburg. http://www.cbs.nl/nr/rdonlyres/8c6e4c9d-4338-4e32-848b-8d43b9b3242d/0/urbanauditiinetherlands.pdf

REFERENCES ON STATISTICAL DISCLOSURE CONTROL

Bethlehem, J. and J. Pannekoek (1998), "Statistical research activities at Statistics Netherlands", *Research in Official Statistics*, 1, pp. 131-134.

Citteur, C.A.W. and L.C.R.J. Willenborg (1993), "Public use microdata files: current practices at national statistical bureaus", *Journal of Official Statistics*, 9, pp. 783-794.

Elliot, M. (2001), "Advances in data intrusion simulation: a vision for the future of data release", *Statistical Journal of the United Nations Economic Commission for Europe*, 18, pp. 383-391.

Elliot, M. and A. Dale (1999), "Scenarios of attack: the data intruder's perspective on statistical disclosure risk", *Netherlands Official Statistics*, 14, pp. 6-10.

Fienberg, S.E. and U.E. Makov (1998), "Confidentiality, uniqueness and disclosure limitation for categorical data", *Journal of Official Statistics*, 14, pp. 385-397.

Gouweleeuw, J.M., P. Kooiman, L.C.R.J. Willenborg and P.-P. de Wolf (1998), "Post randomisation for Statistical Disclosure Control: theory and implementation", *Journal of Official Statistics*, 14, pp. 463-478.

Groot, A. and C.A.W. Citteur (1997), "Accessibility of business microdata", *Netherlands Official Statistics*, 12, pp. 18-32.

Hout, A. van den and P.G.M. van der Heijden (2002), "Randomised response, Statistical Disclosure Control and misclassification: a review", *International Statistical Review*, 70, pp. 269-288.

Hundepool, A.J. (2001), "Computational aspects of statistical confidentiality: the CASC-project", *Statistical Journal of the United Nations Economic Commission for Europe*, 18, pp. 315-320.

Hundepool, A., A. van de Wetering, R. Ramaswamy, P.P. de Wolf, S. Giessing, M. Fischetti, J.J. Salazar, A. Caprara and J. Castro (2003a), *τ-ARGUS, user's manual, version 2.2*, Voorburg, The Netherlands: Statistics Netherlands.

Hundepool, A., A. van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.P. de Wolf, J. Domingo, V. Torra, R. Brand and S. Giessing (2003b), *μ-ARGUS, user's manual, version 3.2*, Voorburg, The Netherlands: Statistics Netherlands.

Hurkens, C.A.J. and S.R. Tiourine (1998), "Models and methods for the microdata protection problem", *Journal of Official Statistics*, 14, pp. 437-447.

Keller, W.J. and J.A. Bethlehem (1992), "Disclosure protection of microdata: problems and solutions", *Statistica Neerlandica*, 46, pp. 5-19.

Kooiman, P., J.R. Nobel and L.C.R.J. Willenborg (1999), "Statistical data protection at Statistics Netherlands", *Netherlands Official Statistics*, 14, pp. 21-25.

Mokken, R.J., P. Kooiman, J. Pannekoek and L.C.R.J. Willenborg (1992), "Disclosure risks for microdata", *Statistica Neerlandica*, 46, pp. 49-67.

Schulte Nordholt, E. (2001), "Statistical Disclosure Control (SDC) in practice: some examples in official statistics of Statistics Netherlands", *Statistical Journal of the United Nations Economic Commission for Europe*, 18, pp. 321-328.

Skinner, C.J. and D.J. Holmes (1998), "Estimating the re-identification risk per record in microdata", *Journal of Official Statistics*, 14, pp. 361-372.

Waal, T. de and L.C.R.J Willenborg (1998), "Optimal local suppression in microdata", *Journal of Official Statistics*, 14, pp. 421-435.

Willenborg, L.C.R.J. (1993), "Discussion statistical disclosure limitation", *Journal of Official Statistics*, 9, pp. 469-474.

Willenborg, L.C.R.J. and T. de Waal (1996), *Statistical Disclosure Control in practice, Lecture Notes in Statistics 111*, New York: Springer-Verlag.

Willenborg, L.C.R.J. and T. de Waal (2001), *Elements of Statistical Disclosure Control, Lecture Notes in Statistics 155*, New York: Springer-Verlag.

Erakunde autonomiaduna
Organismo Autónomo del

EUSKO JAURLARITZA
GOBIERNO VASCO

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA