Netherlands Official Statistics

Volume 15, Autumn 2000

Publisher
Statistics Netherlands
R.L. Vellekoop
Prinses Beatrixlaan 428
2273 XZ Voorburg

Printed by
Statistics Netherlands
Facility Services

Information
Phone: # (31) 45 570 70 70
Fax: # (31) 45 572 62 68
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
www.cbs.nl

Prices do not include postage and
administration costs.
Annual subscription: NLG 42.00 (€ 19.06)
per year
Price per copy: NLG 20.00 (€ 9.08)
Key-figure: A-125/2000
ISSN 0920-2048

Statistics Netherlands

Contents

Knowledge: expenditures, output and growth; provisional results of a knowledge module supplementing the Dutch national accounts

André van den Berg

Abstract

The explanation of economic growth has been given a new impulse by the introduction of the theory of endogenous growth. This theory explains economic growth by, among other things, the growth of productivity, particularly that originating from technological change. Knowledge is a key factor in this theory, and other recent economic literature too recognises the important role of the knowledge present in a society. In order to show explicitly the relationship between knowledge and economics, Statistics Netherlands has developed a so-called knowledge module, supplementing the national accounts.

Keywords: national accounts, knowledge, research and development

1. Introduction

The knowledge module is intended to be a structured set of data on knowledge, linked to the system of national accounts. The conceptual R&D module (Bos et al. 1994) was taken as a starting-point. As the national accounts do not specify most knowledge-related revenues and expenditures as such, a satellite account has to be made to show the relation between knowledge and economics. At the same time the classifications in the standard national accounts, for example those of products and industries, are based on more general economic criteria and thus conceal a lot of information with regard to knowledge. Therefore the standard classifications and concepts were slightly changed in the knowledge module, primarily in the area of output, capital formation and intermediate consumption.

As the development of the knowledge module was started in 1997, the system of national accounts valid at that time must be considered as the reference point in the present article. The national accounts were revised recently and are now in line with the SNA 1993 (UN, 1993) and the ESA 1995 (Eurostat, 1996). Therefore some of the modifications within the framework of the knowledge module which were not standard practice in 1997, have become so in the meantime. The module, as implemented, refers to 1993 and 1994.

2. Aspects of knowledge in the module

According to the theory of endogenous growth, productivity growth is primarily caused by technological changes and innovation. The following aspects of knowledge and of the formation and diffusion of knowledge that can be seen as more or less innovative, are distinguished:

1. Research and development (R&D);
2. 'Soft technology';
3. Education;
4. Incorporated knowledge.

These knowledge aspects are split into an internal and an external part. Expenditures on internal knowledge relate to the own account production of knowledge, while expenditures on external knowledge relate to knowledge bought from third parties. The following expenditures on knowledge are accounted for:

Internal R&D (RD^I):
– compensation of employees involved in internal R&D;
– intermediate consumption related to internal R&D;
– consumption of fixed capital related to internal R&D.

External R&D (RD^E):
– R&D bought from third parties.

'Internal' soft technology and intangibles (ST^I):
– expenditures related to applications for patents, registrations of trademarks and designs.

'External' soft technology and intangibles (ST^E):
– costs of advertising, economic consultancy, patents, copyrights and expenditures on royalties and licences.

Internal education and training (ET^I):
– compensation of employees attending internal company courses;
– compensation of employees organising and teaching at internal company courses;
– intermediate consumption related to internal company courses.

External education and training (ET^E):
– subsidised education;
– compensation of employees attending external company courses;
– external training bought from third parties.

Internal information and computer technology (IT^I):
– compensation of employees involved in internal software development;
– intermediate consumption related to internal software development.

External information and computer technology: (IT^E):
– computers bought from third parties;
– software bought from third parties (computer services excluded).

3. Expenditures on knowledge by industry

Based on the expenditures on the various aspects of knowledge, the industries in manufacturing could be divided into two clusters that could be labelled (A) *innovators* and (B) *followers*. This classification has strong similarities with the clustering into three groups, *high-tech* (H), *medium-tech* (M) and *low-tech* (L) industries as published in *Kennis en economie 1997* (Knowledge and economics, 1997; Statistics Netherlands 1997). The relationship between these two sub-divisions of manufacturing is presented in Table 1. For reasons of statistical co-ordination a split into three was chosen for the presentation of the main findings in Table 2. As expenditures on knowledge may vary strongly from year to year, Table 2 contains the two-year averages of knowledge-linked expenditures.

In 1993 and 1994 over 49 billion guilders on average was spent on knowledge, the equivalent of 4.5 percent of total output (see Table 2). Over one third of this amount involved expenditures on advertising, economic consultancy, patents, copyrights and expenditures on royalties and licences (ST^E). Agriculture, fishing and the low-tech industries in particular spend relatively much on this knowledge aspect (see the middle part of Table 2). In the case of agriculture this may relate to royalties paid for applying cultivation technology under licence. The sectors public services, financial institutions, public administration, subsidised education and health and social work activities spend relatively little on soft technology and intangibles.

Not surprisingly, the high-tech industry in particular can be considered to be knowledge-intensive (see lower part of Table 2). In this industry the expenditures on knowledge account for one sixth of total output. Moreover, all knowledge related expenditures in this industry exceed the macro averages.

The high-tech industry is followed at a distance by the medium-tech industry, financial institutions and public administration, subsidised education and health and social work activities. However, related to output, these industries also spend more on knowledge than average. Agriculture and fishing, mining, quarrying and construction, and electricity, gas and water supply may be considered relatively knowledge-extensive.

Expenditures on R&D (RD^I + RD^E) amounted to almost 12 billion guilders, almost a quarter of total expenditures on knowledge. In both absolute and relative terms R&D is concentrated in the high and medium-tech industries, subsidised education and health, especially in universities and teaching hospitals. As far as external R&D is concerned public administration, etc. are ahead.

Expenditures on information and computer technology (IT^I + IT^E) added up to 11 billion guilders, slightly less than expenditures on R&D. In relative terms financial institutions in particular spent a lot on information and computer technology: 4 percent of total output and nearly 58 percent of total expenditures on knowledge. In absolute terms particularly repair, trade, hotels and restaurants, financial institutions and other business and non-business services (including public administration) are bulk consumers of computers and software (IT^E): 6.1 billion guilders in total.

In relative terms transport, storage and communication spends most on education and training (ET^I + ET^E): nearly one third of their total expenditures on knowledge. The sectors public administration, subsidised education, health and social work activities and electricity, gas and water supply also spend considerable amounts on education. On average, nearly forty percent of all training (in monetary terms) is bought from third parties. Over sixty percent of all training is provided by the organisation itself.

As stated above agriculture and fishing, mining, quarrying and construction and electricity, gas and water supply are relatively knowledge-extensive industries. Although this may be true when expenditures on knowledge are considered in relation to output, a slightly different picture arises, if total compensation of employees is taken as a reference (see Figure 1). From this point of view agriculture and fishing, low-tech industries and also real estate services, renting of movables and business services score above average, while public administration, subsidised education and health and social work score below average.

When we consider expenditures related to knowledge as a fraction of output simultaneously with the same expenditures as a fraction of compensation of employees, we can distinguish three groups:
1. High and medium-tech industries, financial institutions;
2. Agriculture and fishing, low-tech industries, real estate services, renting of movables, other business services, public administration, subsidised education and health and social work;
3. Mining and quarrying, construction, electricity, gas and water supply, repair, trade, hotels and restaurants, transport, storage and communication.

The first group consists of industries that for both ratios score above average. The second group consists of industries that are knowledge-intensive according to only one ratio. Industries in the third group may be considered to be knowledge extensive.

4. Registration of knowledge aspects in the module

The way in which knowledge is registered in the module affects some core macro-economic variables because of the alternative concepts of output, capital formation and intermediate consumption applied in the module. This relates to:
– registration of internal and external R&D as capital formation;
– registration of internal and external software as capital formation;
– introduction of an additional industry households which provides copyrights to non-financial corporations;
– transfer of flows to/from the rest of the world for copyrights, royalties and licenses from the income account to the goods and services account.

In this respect the national accounts data for the years under review (1993 and 1994) serve as a reference. At the moment of writing, national accounts data according to the ESA 1995 revision were not yet available for these years.

The alternative registration introduced in the knowledge module affects the estimates for the core macro-economic variables. Changing the concept of output in the module increases output in 1993 and 1994 by about 17 billion guilders. In both years over 45 percent of this change is caused by the output of internal R&D. On average, output of intangibles contributes to 26.5 percent, while output of internal training contributes to over 16 percent of this change. The remainder, about 12 percent, is accounted for by internal information and computer technology.

In the module a number of knowledge aspects are registered as capital formation while in the national accounts they are registered as intermediate consumption. Altering the concept of capital formation increases capital formation in 1993 and 1994 by about 17 billion guilders, for both years this is an increase in investment by about 15 percent on the national accounts before the ESA 1995 revision. Approximately two thirds of this can be attributed to internal and external R&D. The remainder is caused by the expenditures on internal and external information and computer technology.

On balance the wider concepts of output and capital formation result in an increase in gross value added in 1993 and 1994 by nearly 15 billion and over 16 billion guilders respectively. For both years this equals an increase of 2.9 percent. The adjustments are credited and/or debited to operating surplus.

5. Conclusions

In the modern growth theory, technological change and innovation are an important driving force behind economic expansion. According to this theory, knowledge is an important production factor. The purpose of the knowledge module supplementing the national accounts is to show the different knowledge aspects, viz. expenditures on R&D, education and training, soft technology and information and computer technology. The latter aspect may be seen as an important representative of incorporated knowledge.

In contrast to standard national accounts conventions, expenditures on R&D are not registered as intermediate consumption but as capital formation. This increases investments by about 10 percent

compared with the starting point. In the module as well as in the standard national accounts according to the ESR 1995 conventions, expenditures on information and computer technology, including own-account production of software, is registered as capital formation. Computer-related investment accounts for about 5% of total capital formation.

In the module, the alternative registration increases gross value added by about 2.9 percent. About one percent-point is accounted for by the registration of information and computer technology in accordance with the ESA 1995 conventions.

By comparing expenditures on knowledge with total output by industry and with total compensation of employees by industry, sectors were labelled knowledge intensive or knowledge extensive. This led to a breakdown into three clusters. Medium and high-tech industries and financial institutions are knowledge intensive in every respect, the former because of their relatively high expenditures on R&D and soft technology, the latter because of their relatively high expenditures on information and computer technology. Mining, quarrying and construction, electricity, gas and water supply, repair, trade, hotels and restaurants, transport storage and communication and business services (excluding financial institutions) are knowledge extensive in every respect. No statement can be made about the knowledge intensity of the other industries.

References

Berg, A.T. van den, 2000, *Kennis; uitgaven, productie en groei. Een proefinvulling van de kennismodule* (knowledge; expenditures, output and growth; a pilot for the knowledge module) Statistics Netherlands, Voorburg.

Bos, F.H. Hollanders and S. Keuning, 1994. A research and development module supplementing the national accounts. In: *Review of Income and Wealth*, Series 40, Number 3.

Eurostat, 1996, *European System of Accounts 1995*, Brussel.

Statistics Netherlands, 1992, *Standaard Bedrijfsindeling 1993*, Voorburg.

Statistics Netherlands, 1997, *Kennis en Economie 1977*, Voorburg.

UN, 1993, System of National Accounts, ST/ESA/STAT/SER.F/2/ Rev. 4 United Nations, Washington.

Table 1
Industries [1] per cluster

Cluster A	H/M	Cluster B	M/L
Agro-chemical products, etc.	M	Food products, beverages and tobacco	L
Paint, varnishes, printing ink and mastics, etc.	M	Textiles, wearing apparel, leather and leather products	L
Pharmaceuticals, medicinal chemicals, etc.	H	Wood and products of wood and cork, etc.	L
Soap, detergents, cleaning preparations, etc.	M	Pulp, paper and paper products	L
Other chemical products	M	Publishing, printing and reproduction of recorded media	L
Man-made fibres	M	Coke, refined petroleum products and nuclear fuel	M
Machinery and equipment	M	Basic chemicals	M
Office machinery and computers	H	Rubber and plastic products	M
Electrical machinery and apparatus n.e.c.	H	Other non-metallic mineral products	L
Radio, television and communication equipment, etc.	H	Basic metals and fabricated metal products	L
Medical, precision, and optical instruments, etc.	H	Fabricated metal products (except machinery and equipment)	L
Aircraft and spacecraft	M	Motor vehicles, trailers and semi-trailers	M
Transport equipment n.e.c.	M	Building and repairing of ships and boats	M
		Furniture, manufacturing n.e.c.	M

[1] Classification according to the Dutch industrial classification SBI 1993 (Statistics Netherlands, 1992).
Cluster A: 'innovators'; cluster B: 'followers'.
H: high-tech industries;
M: medium-tech industries;
L: low-tech industries.

Table 2
Expenditures on knowledge aspects; annual averages for 1993–1994 [1] (continued on next page)

	Agri-culture, forestry and fishing	Mining and quarrying and construc-tion	High-tech industries	Medium-tech industries	Low-tech industries	Electricity, gas and water supply	Repair, trade, hotels and restau-rants	Transport, storage and communi-cation	Financial institu-tions	Real estate services, renting of movables and business services	Public admini-stration, subsidised education and health and social work	Other	Total
	billion NLG												
RD^{I}	117	241	1,718	2,724	665	10	161	174	11	267	2,350	0	8,435
RD^{E}	14	55	351	582	372	55	89	30	33	110	1,726	11	3,425
ET^{I}	9	208	246	419	371	76	428	729	562	460	1,798	94	5,400
ET^{E}	26	113	141	286	237	38	287	357	275	268	1,173	96	3,299
IT^{I}	1	20	99	174	143	37	215	119	453	392	293	58	2,004
IT^{E}	65	242	321	633	746	190	1,127	706	1,433	1,723	1,821	328	9,334
ST^{I}	51	69	130	187	66	3	0	58	0	0	0	19	583
ST^{E}	571	931	1,691	2,955	3,048	86	2,532	1,133	496	2,445	384	484	16,756
Total	853	1,877	4,697	7,961	5,648	492	4,837	3,306	3,262	5,665	9,546	1,090	49,235

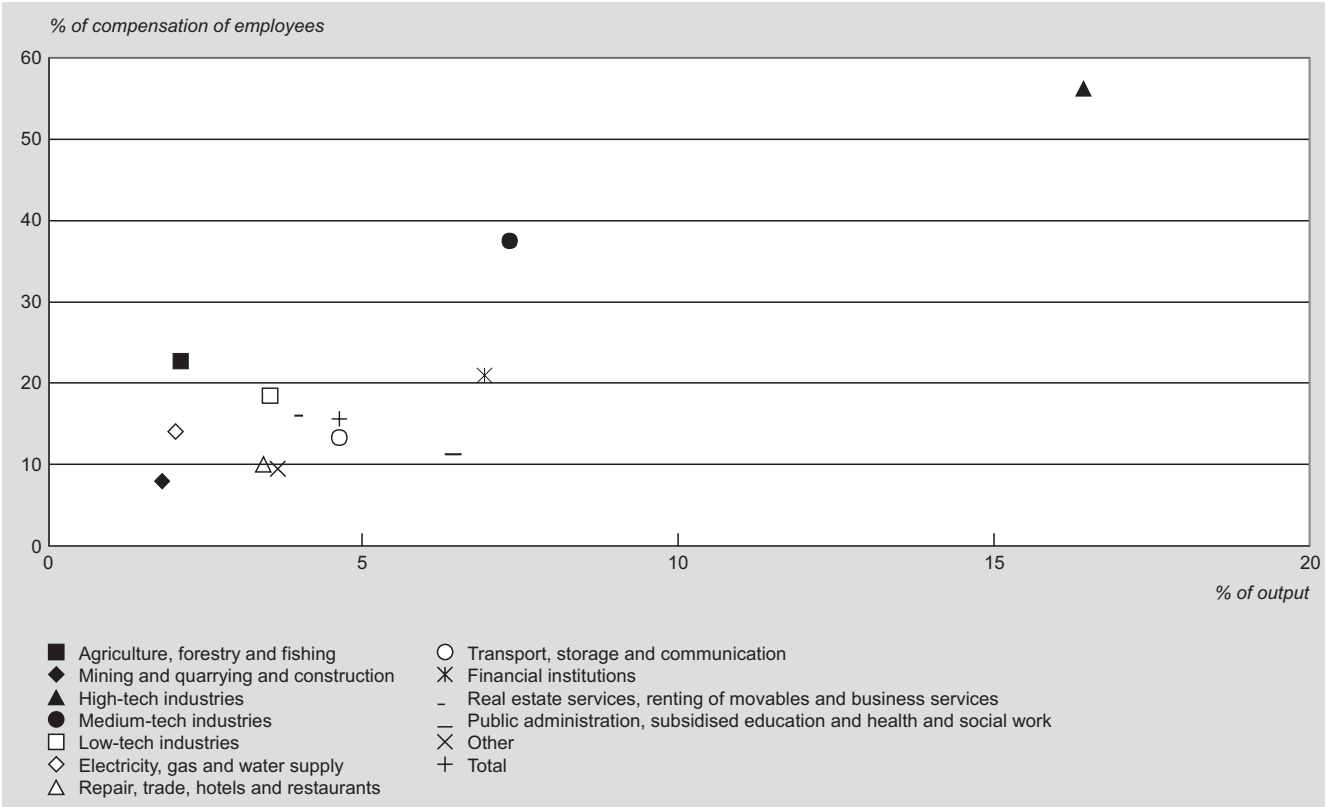
Table 2 (end)
Expenditures on knowledge aspects; annual averages for 1993–1994 [1] (continued on next page)

	Agri-culture, forestry and fishing	Mining and quarrying and construc-tion	High-tech industries	Medium-tech industries	Low-tech industries	Electricity, gas and water supply	Repair, trade, hotels and restau-rants	Transport, storage and communi-cation	Financial institu-tions	Real estate services, renting of movables and business services	Public admini-stration, subsidised education and health and social work	Other	Total
	% of total expenditures on knowledge by industry												
RD^I	13.7	12.8	36.6	34.2	11.8	1.9	3.3	5.2	0.3	4.7	24.6	0.0	17.1
RD^E	1.6	2.9	7.5	7.3	6.6	11.1	1.8	0.9	1.0	1.9	18.1	1.0	7.0
ET^I	1.1	11.1	5.2	5.3	6.6	15.5	8.8	22.1	17.2	8.1	18.8	8.6	11.0
ET^E	3.0	6.0	3.0	3.6	4.2	7.7	5.9	10.8	8.4	4.7	12.3	8.8	6.7
IT^I	0.1	1.1	2.1	2.2	2.5	7.4	4.4	3.6	13.9	6.9	3.1	5.3	4.1
IT^E	7.6	12.9	6.8	8.0	13.2	38.5	23.3	21.3	43.9	30.4	19.1	30.1	19.0
ST^I	6	3.7	2.8	2.3	1.2	0.6	0.0	1.8	0.0	0.0	0.0	1.7	1.2
ST^E	66.9	49.6	36.0	37.1	54.0	17.4	52.3	34.3	15.2	43.2	4.0	44.4	34.0
Total	100	100	100	100	100	100	100	100	100	100	100	100	100
	% output by industry												
RD^I	0.3	0.2	6.0	2.5	0.4	0.0	0.1	0.2	0.0	0.2	1.6	0.0	0.8
RD^E	0.0	0.1	1.2	0.5	0.2	0.2	0.1	0.0	0.1	0.1	1.1	0.0	0.3
ET^I	0.0	0.2	0.9	0.4	0.2	0.3	0.3	1.0	1.2	0.3	1.2	0.3	0.5
ET^E	0.1	0.1	0.5	0.3	0.1	0.2	0.2	0.5	0.6	0.2	0.8	0.3	0.3
IT^I	0.0	0.0	0.3	0.2	0.1	0.1	0.1	0.2	1.0	0.3	0.2	0.2	0.2
IT^E	0.2	0.2	1.1	0.6	0.5	0.8	0.8	1.0	3.0	1.2	1.2	1.1	0.9
ST^I	0.1	0.1	0.5	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1
ST^E	1.4	0.9	5.9	2.7	1.9	0.3	1.8	1.6	1.0	1.7	0.3	1.6	1.6
Total	2.1	1.8	16.4	7.3	3.5	2.0	3.4	4.6	6.9	3.9	6.4	3.6	4.6

Source: Van den Berg, 2000.

[1] Detailed figures are available on request.
RD^I : internal R&D; RD^E: external R&D.
ET^I : internal education and training; ET^E: external education and training.
IT^I : internal information and computer technology; IT^E: external information and computer technology.
ST^I : 'internal' soft technology and intangibles; ST^E: 'external' soft technology and intangibles.

Figure 1. Expenditures on knowledge



Social statistics on the electronic highway; co-ordinated statistics in StatLine

Anne Katrien Amse

Abstract

In 1999 Statistics Netherlands started the project Data matrices on living conditions in StatLine. *The main aim of the project is to publish all information about the living conditions of population groups in a co-ordinated way in StatLine, the electronic output database of Statistics Netherlands. In addition to the basic population figures, living conditions comprise the housing situation, care for household and children, time use, leisure, health, crime and legal protection, social participation, education, labour, income and spending. In short, all statistics produced by Statistics Netherlands which contain information on individuals serve as a source for these co-ordinated publications. The project results in a series of data matrices in StatLine, from which detailed and coherent information can be derived about population groups such as young people and the elderly, immigrants and the native population, working and non-working people, the urban and rural population, and about combinations of these and other features, for all the topics on which Statistics Netherlands produces statistics. The first matrices from these series are by now accessible via the Internet (http://statline.cbs.nl).*

Key words: dissemination of statistical information, statistical co-ordination, publishing, StatLine, social statistics, Internet.

1. Introduction

Several developments have influenced the revision of the publishing process at Statistics Netherlands. The most important of these are a more prominent role for users in determining the output on the one hand, and the possibilities new media offer to make statistical information more accessible on the other.
A few years ago Statistics Netherlands examined the demand for statistical information of several user groups and how it met this demand. The conclusion was that Statistics Netherlands has a good reputation as supplier of reliable *figures*, but that in the field of publishing complete, coherent statistical *information* about branches, population groups and themes like the environment and technology there is room for improvement. Anyone who wanted a complete overview of all information on, for example, the youth in the Netherlands, had to collect these data from dozens of different electronic and/or printed publications. Not only is this very time-consuming, but users were also confronted by the use of different definitions and classifications – different age classes for example – which made the figures incomparable [1)].
To improve this, it was necessary to change from the bottom-up process, in which every statistician determines the output for his own data source, to a top-down process: on the basis of the demand from important user groups, Statistics Netherlands should produce coherent, comparable and complete information, using all appropriate statistical data sources.

New technology made it possible to connect databases at micro level thus enriching them and creating new facilities for analyses and publishing [2)]. On the other hand Statistics Netherlands developed the software for StatLine, the output database of Statistics Netherlands which is accessible via the Internet [3)]. This database makes detailed and co-ordinated publications electronically accessible for the public. In the future these technological developments will act together, when StatBase, a database containing all publishable aggregates, will be filled directly from the micro-databases, to which StatLine provides access through a search and select system.

This was the backdrop for the project *Data matrices on living conditions in StatLine.* The target of this project is the publication in StatLine of all information concerning population groups by several co-ordinated features. It intends to facilitate the selection of many detailed and coherent data about housing, education, labour, health, leisure, crime and legal protection, social participation etc. for groups in the Dutch population such as the young and the elderly, the employed and the unemployed, immigrants and natives, and on combinations of these and other features. As the matrices use standard classifications, the figures are comparable. The project includes all social statistics as well as the business and institution statistics where they concern individuals, for example prison and police statistics. Four of the intended StatLine matrices [4)] are now accessible via the Internet.

2. StatLine: Statistics Netherlands' publication medium

StatLine is the electronic database, in which all publicly available Statistics Netherlands data will be recorded in future. Of course tables and texts which are published in print could be recorded in StatLine in the same way, but the drawback of this is that the limitations of paper – standard A4 format and two dimensionality – are taken as the starting point, without taking the advantages of the electronic medium. The great benefit of StatLine is the possibility to publish large multidimensional data matrices. Such a data matrix can be pictured as a cube in which thousands of cells with co-ordinated data are recorded.
Selections can be made from this cube by choosing at least one item from every selection dimension. One selection dimension is reserved for the subjects: it in fact consists of 'column entries' of the matrix and can contain thousands of variables. The other selection dimensions together form the 'row entries' of the matrix, each comprising a certain classification, for example, one for sex, one for age class, and one for time. Users can select one or more items from each of these classifications and from the subjects to generate a table. The information in a matrix is fully co-ordinated. All subjects are recorded by the same standard classifications, or combinations thereof, so that all tables derived from the matrix comprise comparable information. So, crime figures originating from different data sources, such as police statistics, justice statistics, prison statistics and the surveys on offenders and victims, can be produced by the same combination of classifications. For example: all variables from these data sources on 12–17 year-old boys and girls. All figures can be related to those on the whole population, as these totals are recorded in the matrix as well.
Of course the sources from which the publication figures are derived must contain the variables age and sex, to fill all the cells of the matrix. If part of this information is missing (e.g. only sex and no age, or sex and only some age classes) empty cells emerge in the matrix.

The matrix publications in StatLine have the following advantages:
– Coherence and detail.
  There is more coherence and detail in the published information. Users can select a complete information set about certain population groups such as elderly men and women, immigrant and native youths etc. for all the topics on which Statistics Netherlands produces statistics.

– Accessibility.
  StatLine offers a search option and multidimensional selection possibilities which are not available in printed publications.
– Availability.
  Anyone with an Internet connection can select the desired information from anywhere in the world, at any time.
– More efficient production of printed publications.
  Many current paper publications comprising almost only tables will become redundant. Printed publications with texts explaining and demonstrating relationships between figures, supported by illustrations, will – if there is enough demand – be able to be compiled faster and more efficiently, as the co-ordinated data needed for these are already available in StatLine.

In spite of the great advantages of StatLine, electronic media also have specific limitations and there is therefore certainly room for improvement.

– Empty cells
  A large matrix always comprises empty cells. Certain figures (crossings) are logically impossible, for example: men with cervical cancer as cause of death. In other cases, figures are lacking because Statistics Netherlands does not always survey the whole population, but only a small population group. Data on drug abuse, for example, are only collected for 12–30 year old people. Sometimes there are problems with disclosure risks: if there is only one or very few 12–17 year old boy(s) convicted of manslaughter, these figures may not be published, because they may be traced to individual persons. Sometimes a survey does not cover enough cases to result in reliable data for all the classifications. In general, figures based on fewer than 50 cases in a survey are not published. And sometimes for a certain year under review not all data sources are available yet. There will always be a first statistic for a new year – usually the population statistics which have January 1st as reference date. For all other subjects the data matrix will still be empty for that year. Only when the last statistics are released will all the empty cells eventually be filled with figures.
  Obviously it is annoying when, offered a magnificent menu, you make your choice from it and the selected table comprises empty cells. These empty cells are the inevitable consequence of bringing as much as possible coherence and detail in the publications in StatLine. Basically it must be possible to develop a technical solution for this problem, a tool in the StatLine software that warns when cells without data are selected. For example, for a certain selection, all combinations that produce empty cells, can be made black or impossible to select. No such tool is available at present. Only after a selection has been made and the table is shown on the screen will the user see that the table comprises a few or no figures at all. For the time being, users of StatLine data matrices will have to experiment a bit by selecting data on different aggregation levels or for different years. In spite of the existence of empty cells, the matrix on living conditions of the youth population, for example, comprises many more data than Statistics Netherlands has ever published before about this population group.
– Meta-information about the statistical data sources is lacking.
  When selecting data from data matrices StatLine users should be able to select electronically both the methodological report and the questionnaire related to the survey from which the data are derived. At the moment the website has no facility for this meta-information.
– Varying classifications
  Some classes of the classifications vary, depending on the subject. ‘Total’, ‘other’ and ‘<25 years’ are difficult categories. ‘Total’ does not always relate to the population as a whole. Some statistics have a cut-off at 12, 15 or 18 years. The same problem applies to the class ‘<25 years’. Sometimes this class comprises the age class 0–24, but in other cases it may include 12–24, 15–24 or 18–24. This problem has been overcome by adding this information to the publication unit and/or to the explanatory notes, when the class does not comprise the whole population (up to 25 years). The category ‘other’ is used often to present a remaining group in a series of variables. In a printed publication the meaning of ‘other’ is mostly determined by the groups presented previously: everything not captured by these classes belongs to ‘other’. Because in an electronic publication the category ‘other’ may be selected without selecting all the previous categories, for every ‘other’ category the explanatory notes define precisely what it comprises.

Most of the above-mentioned limitations are temporary; moreover, they are counterbalanced to a great extent by the advantages of the new publishing policy. Eventually, a large, detailed, coherent and up-to-date supply of information on population groups will be accessible on the internet for everybody free of charge.

3. The ideal data matrix on living conditions

Statistics Netherlands aims to publish all ‘publishable’ data in StatLine. This target is easier expressed, however, than implemented. In theory, Statistics Netherlands could serve the public best by simply publishing all the data it collects at micro level, after control and – if necessary – editing. After all this would offer maximal flexibility for aggregation and analysis purposes. In practice, Statistics Netherlands is not allowed to publish figures on individual persons or businesses, and must therefore produce aggregated figures. By definition, the fact that only aggregated figures may be published, diminishes the application possibilities of a database. Moreover, once classifications are chosen, they put up a barrier for other data that are in themselves ‘publishable’. If Statistics Netherlands were to publish data for every possible age class for which publishable data can be produced, this means that – certainly if in addition to age other background features are also used – the limits of the statistical possibilities are exceeded. By combining figures, they can be reduced to data with so much detail that they are no longer reliable, or even to information about individuals. So the first important step is to make choices concerning classifications and combinations of classifications to be used in publications.
This also means that it is an illusion that, once StatLine has been filled with data, it will never be necessary to return to the microdata for specific questions. But a more detailed, coherent and co-ordinated information supply about the living conditions of population groups will certainly decrease the demand for made-to-measure data in this field. With this in mind the following list of demands was set for the StatLine publication on living conditions.
– The demand for information is the central issue.
  The demand for coherent information about population groups determines the design of the data matrix on living conditions, regardless of how many and which data sources and organisation parts must deliver the data.
– Co-ordination is crucial.
  Nationally or internationally acknowledged classifications will be used by determining the row entries of the matrix. These classifications an combinations of classifications must be utilised consequently for all statistics on individuals.
– The advantages of StatLine must be utilised
  As much use as possible will be made of the applications of the electronic medium: flexibility, several selection dimensions, almost no size limits.
– The one-figure notion
  Every variable appears just once in the matrix, regardless how many sources give values – possibly different values for – this same variable. If there is more than one data source, the one with the best quality will be chosen.
– The matrix must be up-to-date
  One of the selection dimensions is reserved for the year under review. The first year is 1997. Whenever a data source has new

material to offer, this will immediately be added to the data matrix in StatLine. Users must be able to rely on the fact that StatLine always comprises the most recent data. Eventually, detailed time series will be available.

First an ideal StatLine matrix with the name *Data matrix on living conditions* was designed. The most important thing in designing a data matrix for StatLine is the design of the column containing the row entries and the corresponding selection dimensions; the design of row with the column headings is less crucial.

3.1 The row entries

We produced a list of important background features which had to be recorded in the row entries. The most difficult classification is the one for age because of the almost unlimited possibilities and because there are hardly any standards for it. A classification into three main groups was chosen: youth (<25 years), elderly people (55 years and older), and the group in-between (25–54 years). This classification is often used in policy and research in the field of young and elderly people respectively. Internationally, people up to and including 24 years of age are considered youths. For the elderly the age limit of 55 years is common in the Netherlands; from this age, for example, labour participation decreases rapidly. In addition, a lower aggregation classification was produced, breaking down the youth and the elderly in further detail: for the youth age-classes 0–3, 4–11, 12–17 and 18–24 years. These age classes more or less coincide respectively with the pre-school period, primary school, secondary school and young adulthood. Eighteen is the age at which children come of age and receive the right to vote for example. The age-class 25–29 was added for purposes of comparison. The elderly are subdivided into the classes 55–64, 65–74 and 75 years and older. Sixty-five is the official retirement age. The categories 'total' and 15–64 years (labour force) are added to all age classifications for comparison.
For the other classifications there is either only one possibility (sex, year, marital status) or a standard classification [5].

For some data sources it is not possible to publish data on the low aggregation level of the classifications, and it is therefore also necessary to record more aggregated classes in the column entries of the data matrix. Thus, in the classification for origin, in addition to Turkey, Morocco, Suriname, the Netherlands Antilles/Aruba and other non-western countries of origin, a class non-western countries was also created, in which all these countries of origin are put together. The purpose is to use the classifications as much as possible in combination in the column entry of a StatLine matrix.

3.2 The column entries

Unlike the row entries, the determination of the variables to be published in the matrix the column entries is in fact not crucial at all. Even when the matrix is constructed and filled with data – new variables can always be added or changed. Changing the variable 'people overweight' into 'people moderately overweight' and 'people severely overweight' will not affect the other cells in the matrix at all, while a change of classification in the column entries will affect every variable in the matrix. In practice, a complete new matrix will then have to be constructed and filled. Although the design of the the column entries thus has fewer consequences, this does not alter the fact that it should be logically structured to make it easier for users to find their way around the matrix.

The column entries of the matrix must contain all variables to be published. Eleven main subjects were chosen: population, housing conditions, care for household and children, time use, leisure, health, crime and legal protection, social participation, education, labour, income and expenditure. Each of these subject blocks is broken down systematically into a hierarchical structure, with at the lowest level the individual variables. Thus the block 'health' is subdivided into 'health status', 'medical consumption', 'parturition and childbirth' and 'causes of death'. 'Medical consumption' is in turn broken down into 'general practitioner', 'specialist', 'dentist', 'physiotherapist', 'hospital' and 'medicine consumption'. Under 'hospital' we find the individual variables are found: 'people admitted to hospital', 'hospital admissions per 100 persons in the population', 'hospital admissions per 100 people admitted to hospital', 'surgery during hospitalisation'.
StatLine is fully equipped to work with such a hierarchical structure, and it is also possible to search for information with key words. Selections can be made at every level in the subject list: a block like 'health' can be selected as a whole, but it is also possible to select only one or several items from a block and combine these with items from another.

3.3 The selections

StatLine has the facility to create a maximum of five selection dimensions in a data matrix. In order to select items from fewer than all available classification selections, it is necessary to record the category 'total' in every classification except the one for years. A user interested in information per age class, but not by sex or other features can generate this information by selecting the required age classes plus the category 'total' from the other selection dimensions. A StatLine data matrix thus shows the user a maximum of four table entries classified by the row entries, and one by the column entries.

This makes it clear that technically it is not possible to record each of the thirteen classifications shown in the appendix to this article in its own selection dimension. Besides, this would mean that the matrix per year already would comprise more than a billion rows, with obviously mostly empty cells. Reliability and confidentiality regulations prevent the publication of data on for example how many 18–24 year-old single mothers of non-western origin with a university degree in the province of North-Holland actively participate in sports. In short: again choices must be made. Several data matrices must be constructed, each of which follows the list of classifications. Not only are they thus internally consistent, but they also offer mutually comparable figures. Although it is not possible to avoid empty cells in a data matrix, the risk of their occurrence is reduced by diminishing the number of selection dimensions in the column entries. The drawback of this is that some figures that are publishable are also cancelled by this limitation. Solutions may be found for these technical problems sometime in the future, so that the ideal data matrix will as yet be published. For the present the decision has been made to develop a series of practical and – considering the present state of technology – attainable matrices. The list of subjects for the column entries is the basis on which every matrix will be constructed, but this will be adapted depending on the availability of the data.

4. A series of practically applicable data matrices on living conditions

Apart from time, age and sex are generally the most important basic classifications of the population, and it is these three classifications that are chosen as the starting point for every matrix, alongside one of the other classifications. The age classification into main groups is used (<25, 25–54, 55+, 15–64 and total), in addition to sex, years and for example individual social group or country of origin. An exception is the classification 'present education level' that hardly ever relates to older people and thus makes the combination with the age classification pointless.
Two matrices comprise only the classifications for sex, age class and years. One is the youth matrix, which uses the age classification for the youth at a low aggregation level; the other is the matrix for the

elderly and uses the age classification for the elderly at a low aggregation level.
In practice, from this starting point a matrix sometimes still has to be split up into two separate ones. This is the case for the social group matrix. A combination of the four classifications in the matrix column with the row entries would generate too many empty cells. One matrix will record the classifications of sex, years and social group, while in another, an aggregated version of the social group, called labour position, will be recorded in addition to sex, age class and years. By making a definite design for the other matrices that are planned, a split into two matrices may be necessary as well.

This is a provisional list of the present and planned matrices on living conditions:

Data matrix on living conditions	Classifications	Status
Youth	sex x age (low aggregation level youth) x years	Ready [6]
Elderly	sex x age (low aggregation level elderly) x years	Planned
Individual social group	sex x individual social group x years	Ready
Labour position	sex x age x labour position x years	Ready
Urbanisation of municipality of residence	sex x age x urbanisation x years	Ready
Origin	sex x age x origin x years	Planned
Household position	sex x age x household position x years	Planned
Marital status	sex x age x marital status x years	Planned
Completed education level	sex x age x completed education level x years	Planned
Present education level	sex x present education level x years	Planned
Social group of the household	sex x age x social group of the household x years	Planned
Personal income	sex x age x personal income x years	Planned
Household income	sex x age x household income x years	Planned
Region/province	sex x age x region/province x years	Planned

At the moment the data matrices on the living conditions of the youth on individual social group, labour position and urbanisation of municipality of residence are published in StatLine. They comprise more than 1,000 subjects and will grow even further. The other matrices will be constructed and published in subsequent years.

5. How to consult StatLine matrices on the Internet?

Statistics Netherlands' general Internet address is http://www.cbs.nl, but there is also a direct StatLine address: http://Statline.cbs.nl . The StatLine database and the general site of Statistics Netherlands are both accessible free of charge. Data can be selected from StatLine in two ways: by searching with – at present only Dutch – key words, or by selection via the tree diagram. By typing one or more keywords and/or years, for example 'crime youth 1998' [7], a list of StatLine publications appears which contain information about this subject. The matrix on the living conditions of youth appears under the name *Youth by age and sex* [8]. The model table presented is a 'dumb' table with random data from the matrix, which will mostly have no bearing at all on the key words entered. It is recommended that users use the facility to make their own selections in the data matrix. Via the selection menu the matrix on the living conditions of youth can be found via statistical themes, people and society, living conditions of population sub-groups, to the matrix 'youth by age and sex'.
Once the matrix has been reached via the keywords or the tree diagram, the desired subjects can be selected from it by simply clicking on them. At least one item must be selected from all four table entries – 'subjects', 'sex', 'age class' and 'years' [9]. Each tab opens a list of subjects. Subjects preceded by + have more subjects available at a lower level. By clicking on + the next level is opened. The individual variables are found at the lowest level, indicated by ♦. A selection can be made at every level by simply clicking on the subject. The selection is indicated by a change of colour. By clicking on the subject again the selection is undone. In the 'subjects' tab, a click on 'health' will select all variables about health, but it is also possible to select some of the individual variables at the lowest level of the subject 'health' together with some of the variables on, for example, 'leisure'. The selected table appears on the screen and may be exported to Lotus/Excel or another format and be downloaded. An example of a table selected from the data matrix is shown below in Dutch and English.

This table was selected in the following way.
The tab *onderwerpen* (subjects) gives the list of main subjects. Open the next level of *bevolking* (population) and click on *totaal*.
Open the next level of *gezondheid* (health) by clicking on the +. The subjects *gezondheidstoestand* (health status), *gebruik van medische diensten* (medical consumption) and *doodsoorzaken* (causes of death) are shown. Click on the + of *gezondheidstoestand*. Click on the subjects (the words, not the +) *roken* (smoking), and *alcoholgebruik* (alcohol consumption). Thus you select all underlying variables. Click on the + of *drugsgebruik* (drugs use) and click on the subjects *marihuana/hashish in de afgelopen 4 weken* (marihuana/hashish in the last four weeks), *cocaine in de afgelopen 4 weken* (cocaine in the last four weeks), *amfetaminen in de afgelopen 4 weken* (amphetamines in the last four weeks), *XTC in de afgelopen 4 weken.* (XTC in the last four weeks), *heroine in de afgelopen 4 weken* (heroin in the last four weeks), and *paddo's in de afgelopen 4 weken* (mushrooms in the last four weeks).
Go to tab *geslacht* (sex) and click on the word *geslacht*. Thus you select 'total', 'man' and 'woman'.
Go to tab *leeftijdsklassen* (age classes), open the list of age classes and click on *18–24 jaar* and *25–29 jaar*.
Go to tab *jaren* (years), open the list of years and click on 1997 and 1998.
Now the selection is complete. The selection can be made in any order and subsequent additional or other selections are always possible. Click on OK and the table will appear on the screen. If you are not satisfied, you can go back to the selection mode by clicking on the button, make appropriate changes and try again. Once the table fulfils your requirements, it can be downloaded by clicking the 'save' button, choosing the required format from the list (e.g. Excel/Lotus), and clicking on the table name that is produced.

Heavy users of statistical data from Statistics Netherlands can take out a quarterly or monthly subscription to the cd-rom version of StatLine. A separate instalment is also available. There is a free demo version of the cd-rom, but a subsequent subscription will be charged. The advantage of the cd-rom is that it is not dependent on

De leefsituatie van de jeugd, de beroeps- en de gehele bevolking [1)]

	Bevolking [2)]	Gezondheid													
	Totaal [3)]	Gezondheidstoestand													
		Roken					Alcoholgebruik			Drugsgebruik					
		Rokers	Zwaardere rokers [4)]	Zwaardere rokers [5)]	Aantal sigaretten/ shags per dag [6)]	Aantal sigaretten/ shags per dag [7)]	Drinkers	Zwaardere drinkers [8)]	Zwaardere drinkers	Marihuana, hashish in afgelopen 4 weken	Cocaine afgelopen 4 weken	Amfetaminen in afgelopen 4 weken	XTC afgelopen 4 weken	Heroine afgelopen 4 weken	Paddo's in afgelopen 4 weken
	abs.	*%*		*in % van sig./ shagrokers*	*abs.*		*%*		*in % van drinkers*	*%*					
Totaal															
18–24 jaar															
1997	1,396,794	38	11	33	5	16	88	28	32	10	2	1	3	0	2
1998	1,356,901	36	10	32	5	15	89	25	29	.	.	.	.	.	.
25–29 jaar															
1997	1,289,180	40	14	38	6	16	86	18	21	8	2	1	1	0	1
1998	1,271,494	39	12	35	5	15	85	17	20	.	.	.	.	.	.
Man															
18–24 jaar															
1997	708,694	39	11	33	6	16	91	40	44	14	3	2	3	0	3
1998	688,172	40	11	31	5	15	91	42	46	.	.	.	.	.	.
25–29 jaar															
1997	657,393	46	18	43	7	17	93	30	32	13	3	2	2	0	2
1998	646,611	46	14	34	6	15	92	28	31	.	.	.	.	.	.
Vrouw															
18–24 jaar															
1997	688,100	37	11	33	5	15	85	16	19	6	1	1	1	0	1
1998	668,729	32	9	32	4	15	86	10	12	.	.	.	.	.	.
25–29 jaar															
1997	631,787	35	10	32	5	15	78	7	9	3	1	1	0	0	1
1998	624,883	32	11	37	4	14	78	7	8	.	.	.	.	.	.

Bron: CBS.

[1)] De leefsituatie van de Nederlandse jeugd (0–24 jaar), de beroepsbevolking (15–64 jaar) en de gehele bevolking naar leeftijd en geslacht.
Deze matrix is nog in ontwikkeling. Regelmatig zullen nieuwe onderwerpen worden toegevoegd.
U kunt de jeugd van 0–24 jaar in zijn geheel selecteren, maar ook daarbinnen groepen onderscheiden naar leeftijdsklasse en geslacht. Van de gekozen groep(en), bijv. meisjes van 12–17 jaar, kunt u op de onderwerpendimensie de gewenste gegevens selecteren. Deze bevat zeer veel onderwerpen, ondergebracht in de hoofdgroepen Bevolking, Wonen, Tijdbesteding, Huishouden en zorg voor kinderen, Vrije tijd, Gezondheid, Criminaliteit en rechtsbescherming, Maatschappelijke participatie, Onderwijs, Arbeid en Inkomen en bestedingen.
Alle gegevens in deze matrix hebben betrekking op de leefsituatie in particuliere huishoudens.
In de matrix zijn gegevens beschikbaar over de jaren vanaf 1997. Zodra nieuwe gegevens beschikbaar zijn, worden deze toegevoegd.

[2)] Het aantal personen in de Nederlandse bevolking, vastgesteld op basis van de Gemeentelijke Basisadministraties op 1 januari. Alleen de gegevens over de bevolking naar huishoudensposities zijn op basis van een steekproefonderzoek vastgesteld.
[3)] Bevolkingsaantal op 1 januari.
[4)] 20 of meer sigaretten of shagjes per dag.
[5)] 20 of meer sigaretten of shagjes per dag.
[6)] Gemiddeld per persoon in de bevolking.
[7)] Gemiddeld per sigaretten of shag rokende roker.
[8)] Minstens 1 dag per week 6 of meer glazen alcoholhoudende drank.

the Internet and therefore will not be hampered by its shortcomings (e.g. long waiting times and peak time problems). In addition to the StatLine cd-rom, Statistics Netherlands publishes cd-rom's with a selection of StatLine publications for specific target groups, for example comprising only data on agriculture; naturally this is cheaper than the complete StatLine cd-rom. If there is sufficient demand, a cd-rom with the series of data matrices on living conditions may be produced some time in the future.

For information or comments on this contribution, please contact Anne Katrien Amse: aame@cbs.nl

StatLine can be found on the internet at: http://Statline.cbs.nl

The StatLine cd-rom can be ordered from: infoserv@cbs.nl

Notes

[1)] The same applies for users who want a complete overview of all available statistical information on one branch of industry.

[2)] See *Netherlands Official Statistics,* autumn 1996 which concentrates on the Economic Statistics Microdatabase and *Netherlands Official Statistics,* summer 2000 which concentrates on the Social Statistical Database.

[3)] See J.W. Altena *The art of cubism*, 1997.

[4)] A warning for non-Dutch speakers: although StatLine is internationally accessible via the Internet and free of charge, only the opening page is currently available in English. At the

The living conditions of young people, the labour force and the total population [1)]

	Population [2)]	Health													
	Total [3)]	Health status													
		Smoking					Alcohol consumption			Use of drugs					
		Smokers	Heavy smokers [4)]	Heavy smokers [5)]	Number of cigarettes per day [6)]	Number of cigarettes per day [7)]	Drinkers	Heavy drinkers [8)]	Heavy drinkers	Marihuana, hashish in the last 4 weeks	Cocaine in the last 4 weeks	Amphetamines in the last 4 weeks	XTC in the last 4 weeks	Heroin in the last four weeks	Mushrooms in the last 4 weeks
	abs.	*%*		*in % of smokers*	*abs.*		*%*		*in % of drinkers*	*%*					
Total															
18–24 years															
1997	1,396,794	38	11	33	5	16	88	28	32	10	2	1	3	0	2
1998	1,356,901	36	10	32	5	15	89	25	29	.	.	.	.	.	.
25–29 years															
1997	1,289,180	40	14	38	6	16	86	18	21	8	2	1	1	0	1
1998	1,271,494	39	12	35	5	15	85	17	20	.	.	.	.	.	.
Man															
18–24 years															
1997	708,694	39	11	33	6	16	91	40	44	14	3	2	3	0	3
1998	688,172	40	11	31	5	15	91	42	46	.	.	.	.	.	.
25–29 years															
1997	657,393	46	18	43	7	17	93	30	32	13	3	2	2	0	2
1998	646,611	46	14	34	6	15	92	28	31	.	.	.	.	.	.
Woman															
18–24 years															
1997	688,100	37	11	33	5	15	85	16	19	6	1	1	1	0	1
1998	668,729	32	9	32	4	15	86	10	12	.	.	.	.	.	.
25–29 years															
1997	631,787	35	10	32	5	15	78	7	9	3	1	1	0	0	1
1998	624,883	32	11	37	4	14	78	7	8	.	.	.	.	.	.

Source: CBS.

1) The living conditions of young people (0–24 years) in the Netherlands, the labour force (15–64 years) and the total population by age and sex.
This matrix is in development; new subjects will be added regularly.
You can select the 0–24 year old group in total, but also select smaller age classes within this group. Via the 'subjects' tab you can select the desired variables for the chosen group (e.g. 12–17 year-old girls). The 'subjects' tab comprises a lot of variables, categorised into the main groups Population, Housing situation, Time use, Care for household and children, Leisure, Health, Crime and justice, Social participation, Education, Labour and Income and Expenditure.
All data in this matrix refer to the living conditions of private households
The matrix comprises data for the years from 1997 onwards. As soon as new data are available, they are added to the matrix.

2) The number of persons in de Dutch population, based on the Basic Administrations of Municipalities on January 1st. The figures on the population by household position are based on a survey.

3) Population figure on January 1st.

4) 20 or more cigarettes or self-rolled cigarettes per day.

5) 20 or more cigarettes or self-rolled cigarettes per day.

6) Average per person in the population.

7) Average per smoker.

8) At least 6 alcoholic beverages, on at least one day per week.

moment the contents of StatLine are in Dutch only. This also means that the search system in StatLine does not respond to English terms. The appendix to this article lists a number of important Dutch terms with their English equivalents, to enable non-Dutch speakers to get an idea of the content of StatLine by selecting data using the Dutch entries.

5) See appendix for the complete list of classifications.

6) StatLine data matrices that have been completed can be consulted via the Internet. However, 'completed' does not mean that all publishable figures are recorded in the matrix. New variables will be added to the data matrices already available as soon as they have been produced for the required column heading.

7) In Dutch*: criminaliteit jeugd 1998*.

8) In Dutch: *Jeugd naar leeftijd en geslacht.*

9) In Dutch: *onderwerpen, geslacht, leeftijdsklasse* and *jaren.*

Appendix A
Translation of Dutch terms in the classifications of the StatLine data matrices on living conditions

Dutch	English
Jaar	*Year*
1997	1997
1998	1998
etc.	etc.
Geslacht	*Sex*
Totaal	Total
Man	Man
Vrouw	Woman
Leeftijdsklasse 1 (hoog aggregatieniveau)	*Age class 1 (high aggregation level)*
Totaal	Total
< 25 jaar	< 25 years
25-54 jaar	25-54 years
55 jaar en ouder	55 years and older
15-64 jaar (beroepsbevolking)	15-64 years (labour force)
Leeftijdsklasse 2 (jeugd, laag aggregatieniveau)	*Age class 2 (Youth, low aggregation level)*
Totaal	Total
0–3 jaar	0–3 years
4–11 jaar	4–11 years
12–17 jaar	12–17 years
18–24 jaar	18–24 years
25–29 jaar	25–29 years
0–11 jaar	0–11 years
12–24 jaar	12–24 years
0–17 jaar	0–17 years
18–29 jaar	18–29 years
0–24 jaar	0–24 years
15–64 jaar	15–64 years
Leeftijdsklasse 3 (ouderen, laag aggregatie niveau)	*Age class 3 (Elderly, low aggregation level)*
Totaal	Total
55–64 jaar	55–64 years
65–74 jaar	65–74 years
75 jaar en ouder	75 years and older
55 jaar en ouder	55 years and older
65 jaar en ouder	65 years and older
15–64 jaar	15–64 years
Huishoudenspositie	*Household position*
Totaal	Total
Eenpersoonshuishouden	One-person household
Samenwonend paar zonder kinderen	Couple living together without children
Ouder samenwonend paar met kinderen	Parent in household of couple with children
Kind samenwonend paar met kinderen	Child in household of couple with children
Ouder eenouderhuishouden	Parent in one-parent household
Kind eenouderhuishouden	Child in one-parent household
Overig particulier huishouden	Other private household
Institutioneel huishouden	Institutional household
Burgerlijke staat	*Marital status*
Totaal	Total
Nooit gehuwd	Never married
Gehuwd	Married
Geregistreerd partnerschap	Registered cohabitation
Weduwstaat	Widowed
Gescheiden	Divorced
Individuele sociale groep	*Individual social group*
(personen getypeerd naar hun individuele sociale groep)	(persons according to their individual social group)
Totaal	Total
Werkzame beroepsbevolking totaal	Active labour force, total
Hogere leidinggevenden, hooggeschoolde hoofdarbeid	Higher management, highly skilled non-manual employees
Overige hoofdarbeid, MBO-niveau	Other non-manual employees
Kleine zelfstandigen, zelfstandige agrariers	Small proprietors, self-employed farmers

Supervisoren handarbeid, geschoolde handarbeid
Half- en ongeschoolde handarbeid, agrarische arbeid
Werkloze beroepsbevolking
Niet-beroepsbevolking totaal
Onderwijsvolgend
Arbeidsongeschikt
Gepensioneerd
Verzorgt huishouden met minderjarige kinderen
Verzorgt huishouden zonder minderjarige kinderen, en overigen

Supervisors of manual employees, skilled manual employees
Semi- and unskilled manual employees, agricultural employees
Unemployed labour force
Non-labour force, total
In education
Disabled
Retired
Carer in household with minors
Carer in household without minors, and others

Sociale groep huishouden
(personen geclassificeerd naar de sociale groep van het huishouden waar ze deel van uitmaken. Deze wordt gebaseerd op de sociale groep van de persoon in het huishouden die het hoogste inkomen genereert).
De classificatie is gelijk aan die van de individuele sociale groep

Social group of household
(Persons according to the social group of the household to which they belong, based on the household member with the highest income).

Classification as for individual social group.

Besteedbaar persoonlijk inkomen
Totaal
1e kwartiel
2e kwartiel
3e kwartiel
4e kwartiel

Personal disposable income
Total
1st quartile
2nd quartile
3rd quartile
4th quartile

Netto huishoudinkomen
(personen geclassificeerd naar het inkomen van het huishouden waar zij deel van uitmaken)
De classificatie is gelijk aan die van besteedbaar persoonlijk inkomen

Net household income
(persons according to the income class of the household to which they belong)
Classification as for personal disposable income.

Behaald onderwijsniveau
(personen geclassificeerd naar hoogst behaald diploma)
Totaal
Basisonderwijs
Mavo
Vbo/Sbo1
Sbo2
Sbo3
Sbo4
Havo/Vwo
Hbo
Wo
Vbo/sbo/mavo (Voortgezet onderwijs, lager)
Havo/vwo/Sbo2,3,4 (Voortgezet onderwijs, hoger)
Hbo/wo

Completed level of education
(Persons according to their highest diploma)
Total
Primary school
Junior general secondary education
Pre-vocational education
Basic vocational training (level 2)
Professional training (level 3)
Middle-management training (level 4)
Senior general secondary education/pre-university education
Higher vocational education
University education
Lower secondary education
Upper secondary education
Tertiary education

Huidig gevolgd onderwijsniveau
(onderwijsvolgende personen getypeerd naar huidig gevolgd onderwijsniveau)
De classificatie is gelijk aan die van het behaalde onderwijsniveau

Present education level
(persons in education according to the education level in which they are active at the moment)
Classification as for completed level of education.

Herkomst
(personen geclassificeerd op basis van het eigen geboorteland en dat van hun ouders. Personen van wie minstens één ouder in het buitenland is geboren, worden beschouwd als allochtoon. Europa (excl. Turkije), Noord-Amerika, Indonesië, Japan en Oceanië worden als westerse landen beschouwd en Afrika, Azië (excl. Indonesië en Japan, Latijns-Amerika en Turkije vallen onder de niet-westerse landen.

Origin
(persons according to their country of birth and that of their parents. People with at least one parent born outside the Netherlands are defined as immigrants.
Europe (excl. Turkey), North America, Indonesia, Japan, and Oceania are considered to be 'western' countries. Africa, Asia (excl. Indonesia and Japan), South America, and Turkey belong to the 'non-western' countries

Naar land van herkomst
Totaal
Nederland
Turkije
Marokko
Suriname
Antillen/Aruba
Overige niet-westerse landen
Westerse landen
Turkije/Marokko

By country of origin
Total
Netherlands
Turkey
Morocco
Suriname
Netherlands Antilles/Aruba
Other non-western countries
Western countries
Turkey/Morocco

Suriname/Antillen, Aruba	Suriname/Netherlands Antilles, Aruba
Turks/Marokkaans/Surinaams/Antilliaans/Arubaans	Turkish/Moroccan/Surinamese/Antillean/Aruban
Totaal niet-westerse allochtonen	Non-western immigrants, total
Totaal allochtonen	Immigrants, total
Naar generatie	*By generation*
1e generatie (zelf geboren in buitenland)	1st generation (not born in the Netherlands)
2e generatie (zelf geboren in Nederland)	2nd generation (born in the Netherlands)
Stedelijkheid woongemeente	*Urbanisation degree of municipality of residence*
(Classificatie gebaseerd op gemiddelde omgevingsadressendichtheid)	(Classification based on average surrounding address density)
Totaal	Total
Zeer sterk stedelijk	Very strongly urbanised
Sterk stedelijk	Strongly urbanised
Matig stedelijk	Moderately urbanised
Weinig stedelijk	Hardly urbanised
Niet stedelijk	Not urban
Provincie	*Province*
Totaal	Total
Groningen	Groningen
Friesland	Friesland
Drenthe	Drenthe
Overijssel	Overijssel
Flevoland	Flevoland
Gelderland	Gelderland
Utrecht	Utrecht
Noord-Holland	North Holland
Zuid-Holland	South Holland
Zeeland	Zeeland
Noord-Brabant	North Brabant
Limburg	Limburg
Noord	North
Zuid	South
Oost	East
West	West

Determinants of days lost from work because of illness

Martin Boon

Abstract

This paper investigates the impact of a number of company variables, such as sector of economic activity, labour productivity, capital intensity and market share, on sick leave in the Dutch private sector. Controls are included for the effect of personal and job characteristics. The regression analysis is based on individual company-level data from the quarterly sick leave survey, the annual production statistics, the annual employment and wages survey and the registration system of the social security funds for (the fourth quarter of) the year 1996. The estimation results showed that the company's size, the sector of economic activity, the age of the workers, the working schedules (regular/irregular/shift work), the mean hourly wage and the working hours have a significant influence on the number of days lost from work due to illness. For other personal, job and company characteristics (such as labour productivity) we do not find a significant effect on sick leave.

Keywords: sick leave, explanation, company, personal and job characteristics

1. Introduction

The number of disabled workers is in the Netherlands rather high compared with neighbouring countries: 12.7 percent of the labour force received a disability benefit in 1996. Generally speaking, workers will be absent from work before they become officially occupationally disable, and therefore a reduction in absence will lead to a reduction of people receiving benefits under the Occupational Disability Act (WAO). Before measures can be taken against absence caused by illness, research has to be undertaken into its determinants. In other words, we have to look at the personal, job and company characteristics that affect the level of sick leave. As sick leave indicates the number of days lost from work because of illness, this research may also help reduce the costs incurred by companies because of illness of their employees.
The economic literature on sick leave concentrates mainly on the supply side of the labour market. Involuntary absence occurs if employees have to work more hours than is desirable. From this point of view particularly the effect on sick leave of personal characteristics, such as sex and age, and job characteristics are considered. Although the demand side of the labour market is usually underexposed, the possible role of employers should not be ignored: employers can really influence absentee rates. For this reason, this paper examines the effect on sick leave of the characteristics of the companies for which people work. It is conceivable, for instance, that variables like sector of economic activity, labour productivity, capital intensity and market share can affect sick leave. The analysis includes controls for the effect of supply side factors, i.e. personal and job characteristics.
Until 1996 the most complete source for information on sick leave was the Health Insurance Act, under which the Board of Supervisors of Social Security Funds (CTSV) produced statistics based on registrations of the industrial insurance boards. However, on 1 March 1996 the Health Insurance Act was privatised and employers became responsible for the payment of (at least 70 percent of) the wage during the first year of their employees' absence. As a consequence of this measure, the data on the nature and duration of worker absence are no longer systematically registered by the industrial insurance boards. Now, the Health Insurance Act statistics compiled by the CTSV only consist of absence information for specific population groups, for example workers with a flexible employment contract, for whom employers are not obliged to pay wages in the case of illness.
The quarterly sick leave survey of Statistics Netherlands provides data on absence caused by illness and is conducted among a sample of companies. This survey only gives information about the effect of sector of economic activity and company size on sick leave. In addition, the Labour Force Survey compiles information on the absence of individual employees among a sample of workers. This employee-level data can help to establish the effect on absence of personal characteristics such as sex, age, marital status, education level and occupation, but not of company characteristics (with the exception of sector of economic activity). By means of linking company level data from different surveys, we can determine the effect of both company characteristics and personal or job characteristics on absence caused by illness. In this respect we consider the quarterly sick leave survey, the annual production statistics, the annual employment and wages survey and the registration system of the social security funds.
The purpose of the present article is to analyse the effect of a number of company variables, such as labour productivity, on the sick leave in Dutch manufacturing companies. The analysis is based on cross-sectional linked company level data for 1996. Section 2 reviews previous studies of the impact of personal, job and company variables on sickness absence. In section 3 the data used in this study are described. Section 4 presents the empirical results and section 5 contains some concluding remarks.

2. Previous studies on sick leave: effects of personal, job and company variables

There are a number of hypotheses about the possible relationship between specific variables and the labour absence rate. This section gives an overview of previous studies that have tested these hypotheses empirically. We distinguish between the effects of personal, job and company characteristics. See also Smulders (1984a,b) for an extended picture of the variables that affect sickness absence.

2.1 Personal characteristics

The demographic composition of a company's workforce plays a role in occurrence of sickness absence. Using information of individual employees Otten and Smeets (1995) investigated the impact of personal characteristics, such as *sex*, on the probability of worker absence. They found that women have a higher chance of staying home sick than men, mainly because more often than men they tend to take care of other members of the family who are ill.
The *age* composition of the workforce may also have implications for absentee rates. Vistness (1997) and Otten and Schmeets (1995) concluded that age is inversely related to worker absence: younger workers have a higher probability of missing time from work. The explanation usually given for this that younger workers change jobs more frequently and job search might be undertaken off the job through labour absence. The negative correlation between age and absence can also be attributed to greater attachment to work and fewer injuries among older workers than among younger workers. However, other researchers such as Groot (1998) found that worker absence is higher among older employees than younger ones. This can probably be explained by the fact that younger workers are healthier. In other words, age can also be interpreted as a proxy for health status.
In some studies *marital status* emerges as a determinant of worker absence. Vistness (1997) shows a positive correlation between the number of children under the age of six and the likelihood of

absence. This finding can be explained by the fact that parents have to take leave to care for sick children. Groot (1998) found that married workers tend to be absent for longer periods than single workers, confirming the finding that increased family responsibilities may increase absentee rates.

2.2 Job characteristics

Part of absence from work can be attributed to working conditions. Groot (1998) studied the effect of job characteristics, in addition to personal characteristics, on sick leave. He concluded that the quality of the connection between job and capacities of the employee, and the physical and mental burden of the job were the most important determinants of absence from work. Otten and Schmeets (1995) investigated which specific forms of working load affect the duration of sick leave, after controlling for the effect of personal variables. It appeared that among other aspects dangerous and heavy work, little pleasure at work, and insufficient possibilities to develop one's talents are associated with high sickness absence rates.

Generally it is assumed that employees who do *irregular and shift work* have relatively higher absentee rates. Remarkably, though, Groot (1998) concluded that sick leave is precisely lower among employees who work on shift rota. He explains this by so-called self-selection: people who accept shift work are a self-selecting group of employees who are relatively healthy and have a low probability of sickness absence.
In a standard labour supply framework of worker absence that models the trade-off between income and leisure time (Kenyon and Dawkins, 1989), we cannot draw a conclusion about the sign of the relationship between labour absence and *wage* rate. According to the substitution effect , absence decreases with increasing wages, while the income effect shows an increasing absence with increasing wage. Using panel data of individual companies, Kenyon and Dawkins (1989) and Brown, Fakhfakh and Sessions (1999) found a significant negative relationship between wage rate and sickness absence. This result implies that the substitution effect dominates any income effect.
From the point of view of the demand side of the labour market, sick leave can be seen as costly to companies because it results in lower productivity (Coles and Treble, 1996). The production loss through worker absence is particularly high in companies with assembly line production processes, because in such a production process the workers are complementary to each other and if one worker is absent the productivity of the remaining workforce is reduced. Such companies are willing to pay higher wage costs to reduce absence. Thus, in the context of efficiency-wage considerations there is a negative correlation between wages and sick leave.
According to the labour-leisure model there is a positive relationship between labour absence and the number of *working hours*. This result can be explained by the fact that more working hours leads to the availability of less time to recuperate from an illness and, presumably, an incentive to compensate for that reduced time by taking sick leave. For instance, workers with a *part-time* or *flexible* employment contract are expected to have relatively lower absence rates because of their flexible work schedules. In addition, one could conjecture that full-time employees who do too much overtime under high pressure have a higher sick leave rate. This is caused by the well-known burn-out syndrome, which is mainly characterised by a state of emotional and mental exhaustion. The assumed positive relationship between working hours and absence is confirmed by the empirical study of Kenyon and Dawkins (1989). However, Vistness (1997) found a negative influence of work time on labour absence, explained by the fact that people who work for longer hours enjoy their work (i.e. have a higher job satisfaction) and may therefore be less likely to be absent than others.
Tenure within a company has an ambiguous effect on absenteeism. On the one hand, tenure could indicate that a worker is more secure in his job and less worried about the repercussions of frequent absence. In this case, an longer period of employment increases the probability of absence. On the other hand, tenure could be correlated with increased job satisfaction and more pleasant working conditions, which would predict a negative relationship with absence. Vistness (1997) found a significant negative effect of tenure on labour absence, in agreement with the latter hypothesis.

2.3 Company characteristics

The characteristics of employing companies can throw more light on the differences in rates of sick leave. For instance, there may be a correlation between *company size* and absence due to illness. Previous empirical research, such as that by Vistness (1997), showed that the absence rate is lower in small than in large companies. He suggests that workers in larger companies feel a sense of alienation (i.e. are not closely connected with the company) and may be more likely to be absent from work than workers in small companies. In addition, employees of large companies may feel more confident than those of small companies that their colleagues can fill in for them in their absence.
There may also be differences in absence rates between *sectors of economic activity*: some sectors, such as manufacturing, have relatively high accident rates because of the dangerous and heavy work, while others, such as health care, the absence is high because of high work pressure and changing rotas. There are also differences between sectors in the measures taken to prevent or limit absence due to illness. Possible measures are for instance joining an organisation for occupational health and safety, monitoring sick employees and rewarding employees who miss fewer days through illness.
Another hypothesis concerns the positive connection between labour productivity and labour absence. The maximum load-bearing capacity of an employee depends mainly on characteristics such as sex, age and health status. The actual load is affected by factors such as working conditions, required output, fringe benefits, and number of working hours. Workers will report in sick when the actual load exceeds the threshold of the load-bearing capacity. *Labour productivity* can be taken as a proxy for the actual load. Using a regression analysis on macro-data for a number of years Fase and Keijzer (1991) showed that labour productivity has a positive effect on worker absence. A problem is that there is not only a causal relationship from labour productivity to sickness absence but also in the opposite direction, as a high absence rate has negative effects on labour productivity. This implies that labour productivity is an endogenous explanatory variable.

The degree of competition in the product market of the company can affect the absence rate. The working load for employees is higher in companies which operate in highly competitive markets. Such companies need to enhance productivity by means of product and process innovation again and again. The impact of competition on absence is measured by the company's *market share*, its *export share* and the four-company *concentration ratio* (C4). The market share, which is defined as company's sales divided by sales of all companies in the relevant sector, determines the extent to which the company can control the sector's output. As the market share is an inverse measure of competition, we expect its impact on worker absence to be negative. The export share of a company is defined as foreign sales divided by total sales. A high export share, indicating that the company is exposed to high competition on foreign markets, enhances the probability of worker absence. The four-company concentration ratio is defined as employment in the four largest companies divided by total employment in the companies in the relevant sector. A competitive market structure, as indicated by a low market concentration, is expected to be associated with a high absence rate.
The *capital intensity* of the production process of a company can also influence labour absence due to illness: sick leave is likely to be higher in labour intensive companies. The capital intensity can be measured by depreciation costs per worker.

3. Data description

The data used in this study are based on cross-sectional information on individual companies in the Dutch manufacturing sector for (the fourth quarter of) the year 1996. The data were created at Statistics Netherlands by matching microdata from the quarterly sick leave survey, the annual production statistics, the annual employment and wages survey and the registration system of the social security funds.

Since the fourth quarter of 1995 Statistics Netherlands has been conducting a quarterly survey into sickness absence of employees of private companies (Keij-Deerenberg, 1996). This quarterly sick leave survey (KVZ) covers all companies with 100 or more employees and a sample of the smaller ones, drawn from the general business register (ABR). Companies are asked to report the average rate of sickness absence, both including and excluding maternity leave, and the kind of sick leave insurance. Following the standard definition of worker absence by the Netherlands institute for occupational health (NIA) (1996), the rate of sickness absence is defined as the number of days lost from work due to illness, accidents, legal or contractual reasons (maternity leave, death in family) and other reasons (labelled 'voluntary'), as a percentage of the total number of available working days of the employees in the relevant quarter. This percentage indicates the magnitude of worker absence in relation to labour capacity. In reply to the question on kind of sick leave insurance the company has to indicate how the company has arranged the obligation to continue the wage payment during the first absence year of their employees:

– fully at company's own risk (i.e. no reinsurance);
– wage payment partly reinsured;
– wage payment completely reinsured.

We have used the KVZ data for the fourth quarter of 1996. When we look at the sickness absence in the course of time (see Table 1), we see that the absence rate in the chosen quarter does not deviate strongly from the rates in the other quarters.

Table 1
Sickness absence rate (excluding maternity leave), private companies, 1995–1998

Year	Quarter	Absence rate (in %)
1995	4	5.0
1996	1	5.3
	2	4.4
	3	4.1
	4	4.6
1997	1	4.9
	2	4.4
	3	4.1
	4	4.9
1998	1	5.2
	2	4.9
	3	4.5
	4	5.3

In the annual production statistics (PS) companies in the manufacturing sector are asked for detailed information on inputs and outputs. This information contains, among other things, domestic and foreign sales, wage bill, depreciation costs (CBS, 1998). The companies are selected from the ABR. Since 1987 all companies with twenty or more employees are surveyed and from the companies with fewer than 20 employees a sample is drawn.

We have also used the registration system of social security funds (VZA) for the year 1996. The VZA comprises individual information on all employees in the private sector which are insured for the obligatory employee insurance such as that under the industrial disability act (WAO) (Van Toor and Schaafsma-Harteveld, 1997). Thus, government employees are not included in the VZA. In the VZA we can distinguish between information on employment (industrial insurance board, indicator temporary employee/worker on call), wage slip (gross annual wage), employers (registration number with an industrial insurance board), personal characteristics (age, sex, marital status i.e. married/unmarried) and addresses. As a person registered in the VZA may have one or more employers from whom he receives a wage in the relevant year, the VZA contains all insured individuals who worked at a certain company during the relevant year. To combine this information with the KVZ data for the fourth quarter 1996, we used the VZA data on the start and the end of the employment period to check whether an individual was employed at a company on the reference date of 31 December 1996.

Lastly, we used microdata from the annual employment and wages survey (EWL). The EWL provides data on the structure of earnings and employment in companies with one or more employees. The data are obtained partly from salary administrations (Arnoldus, 1997) and partly through written surveys. Some of the sampled companies supply an electronic file with details of all their employees. For each employee, data are provided on some personal characteristics (e.g. age), job characteristics (e.g. tenure; employment contract: full-time/part-time/flexible; work schedules: regular/irregular/shift work), gross monthly wages (excluding overtime), hours worked. The gross monthly wages are converted to gross hourly wages using information on hours worked per year. Alongside electronic data collection, complementary data are obtained using paper questionnaires. In the paper survey, mostly small companies are asked to supply data on a sample of all their employees.

Our cross-sectional data set is the result of linking microdata from the above-mentioned four surveys, i.e. KVZ, VZA, EWL and PS. We examined how representative the linked data set for manufacturing companies is by comparing it with the original data set of the PS for the entire manufacturing sector. Table 2 contains summary statistics of the data sets; it appears that companies in the linked data set are larger than in the PS data set. The larger average company size reflects the sample design for the KVZ and the PS, in other words drawing a sample from the smaller companies. We also find that a large portion of companies in the linked data set are

Table 2
Summary statistics for the original PS data set and the linked data set, manufacturing companies, 1996

	Data set	
	PS	KVZ-VZA-EWL-PS
Mean		
Company size	24	221
Labour productivity [1)]	128	160
Capital intensity [2)]	18	25
Wage costs per employee [3)]	74	85
% of companies in sector of economic activity (3-digit level)		
Food and stimulants	16.9	16.9
Textiles and apparel	6.3	5.3
Leather	0.9	1.0
Wood	4.4	1.8
Paper and publishing	16.6	12.3
Petroleum and metal	15.7	14.3
Chemicals	2.2	9.3
Rubber and plastic	3.2	5.6
Other non-metallic mineral	3.4	4.8
Machinery and equipment	9.3	10.9
Computers and electrical apparatus	7.4	7.3
Transport equipment	4.8	5.6
Furniture and recycling	9.1	4.9
Number of companies	34,019	1,540

1) Value added per employee in thousands of guilders.
2) Depreciation costs per employee in thousands of guilders.
3) In thousands of guilders.

active in the chemical industry, while the paper and publishing industry accounts for a smaller portion.

4. Empirical results

In Section 2 a number of hypotheses concerning the effects of a number of variables on sickness absence were formulated with the help of theoretical and empirical results of previous research. With multiple regression analysis these hypotheses are tested. To determine the proper effect of sex on absence, we take the sickness absence excluding maternity leave as the dependent variable.

4.1 Private companies

Table 3 shows the estimation results of a regression model for the explanation of worker absence based on two data sets for private companies. In the columns the ordinary least squares (OLS) estimates of the regression parameters along with their t-values and statistical significance levels are given. We used the following linked data sets for the year 1996: KVZ-VZA and KVZ-VZA-EWL. The effects of exogenous variables on sick leave are analysed by means of a log-linear regression model that relates the logarithm of sickness absence to the logarithm of the exogenous variables (with the exception of the aggregated dummy variables such as the percentage unmarried). The reason is that after the logarithmic transformation the distribution of the variable sick leave is less skew and approaches the normal distribution better. In case of a logarithmic specification the estimated parameters are elasticities, i.e. they denote the percentage change in absence which results from a one percentage rise in the given explanatory variable.

First, we consider the effects of the personal characteristics on worker absence. The regression model including personal variables and only three company variables (size, sick leave insurance, sector) results in an multiple correlation coefficient (R^2) of 15% (see Table 3; first column). This means that 15% of the variation in the (logarithmic) sickness absence is explained by the given model. In the data sets KVZ-VZA and KVZ-VZA-EWL *age* has a significant positive effect on the absence rate. Sick leave is lower in companies that employ relatively many workers under 35 years of age, probably because younger employees are relatively healthier. In our data sets *sex* has a significant impact but its coefficient has a conflicting sign. Based on the KVZ-VZA data we find that the percentage of male workers within a company has a positive effect on absence due to illness (excluding maternity leave), while the KVZ-VZA-EWL data show a negative effect for sex. This difference can probably be attributed to a selectivity bias introduced by linking the data set with the EWL. The estimation results based on the data sets KVZ-VZA and KVZ-VZA-EWL show that *marital status*, in terms of the percentage of unmarried workers, has no significant effect on worker absence.

Next, we look at the impact of job characteristics on sickness absence. The addition of the job characteristics wage, tenure, employment contract, working schedules and working hours to the sickness absence equation increases the R^2 by 1.3% in the KVZ-VZA-EWL data (compare columns (2) and (3) of Table 3). From the coefficient estimates in Table 3, it can be inferred that the mean hourly wage, the percentage of workers with regular work times and the mean number of working hours have a significant effect on the absence rate. The estimation results based on the data set KVZ-VZA-EWL show a significant negative effect of *mean hourly wage* on the absence rate. This finding agrees with the assumption that in the case of a wage increase the substitution

Table 3
OLS estimates of the logarithmic regression model for sickness absence (excluding maternity leave), private companies, 1996 [1)]

Variable	KVZ-VZA	KVZ-VZA-EWL	
	(1)	(2)	(3)
Constant	2.912 (10.47)**	0.990 (11.35)**	0.277 (0.53)
Personal characteristics			
% unmarried	–0.0003 (0.50)	–0.0001 (0.10)	0.0005 (0.59)
% male	0.001 (2.46)**	–0.002 (3.30)**	–0.002 (2.97)**
% younger than 35 years	–0.004 (7.81)**	–0.004 (5.33)**	–0.005 (7.09)**
Job characteristics			
Mean hourly wage (log)			–0.351 (9.10)**
Mean tenure (log)			0.023 (1.43)
% full-time workers			–0.001 (1.77)
% workers with regular hours			–0.0006 (1.80)*
Mean working hours (log)			0.354 (4.90)**
Company characteristics			
Size (log)	0.143 (24.25)**	0.155 (21.36)**	0.154 (20.79)**
Sick leave (re)insurance	–0.048 (2.71)**	–0.009 (0.51)	–0.016 (0.88)
Sector of economic activity (2-digit level): [2)]			
Agriculture and fishing	0.011 (0.19)	–0.034 (0.41)	–0.100 (1.23)
Mining	–0.141 (1.22)	–0.251 (2.12)**	–0.260 (2.21)**
Manufacturing	0.096 (2.35)**	0.103 (2.32)**	0.049 (1.08)
Public utilities	0.162 (1.66)	0.155 (1.66)	0.172 (1.85)*
Construction	0.012 (0.27)	0.041 (0.78)	0.017 (0.33)
Trade	–0.145 (3.53)**	–0.186 (4.07)**	–0.230 (4.99)**
Hotel and catering	–0.088 (1.42)	–0.190 (2.51)**	–0.188 (2.47)**
Transportation and communication	–0.091 (1.84)*	–0.106 (2.01)**	–0.182 (3.42)**
Financial institutions	–0.011 (0.16)	–0.127 (1.83)*	–0.100 (1.43)
Business services	–0.149 (3.52)**	–0.152 (3.33)**	–0.162 (3.55)**
Public government	0.029 (0.25)	–0.059 (0.52)	–0.056 (0.50)
Education	0.107 (0.99)	0.121 (0.95)	0.164 (1.30)
Health care	0.322 (7.26)**	0.213 (4.48)**	0.164 (3.37)**
R^2	0.153	0.167	0.180
Number of companies	9,438	6,655	6,655

1) Absolute values of t-values are mentioned in brackets.
** = significantly different from zero at the 95% level;
* = significantly different from zero at the 90% level.
2) The sector Culture or other services is excluded to avoid multicollinearity.

effect dominates the income effect. As expected, we found that the sickness absence is high in companies with a relatively high percentage of employees working in *shifts or irregular hours*. There appears to be a positive relationship between absence and *number of working hours*. This agrees with the fact that employees who work too much overtime under high pressure have a higher absence rate. Further, we can conclude that there is no significant relationship between labour absence and either the percentage of full-time workers and the mean tenure.
Lastly Table 3 shows how absentee rates vary with a limited number of company characteristics for the private sector. From the estimates we can conclude that company size and the sector of economic activity have a significant effect on worker absence in all three regressions. The absence rate is relatively higher in *larger companies,* corresponding with the fact that workers in these companies do not feel closely connected with the well-being of the company. Absence rates in the trade, transport and communications, and business services *sectors* is relatively low compared with the rest of the private sector. As expected, it appears that there is a relatively high absence rate in health care. For the dummy variable indicating the presence of a sick leave insurance we find that the absence rate is lower in companies that have insured the risk of worker absence.

4.2 Manufacturing companies

In columns (2) and (3) of Table 4 the OLS estimation results based on the KVZ-VZA-EWL-PS data set are shown for manufacturing companies. With this data set we can calculate the absence effect of complementary company variables such as labour productivity. As already mentioned in Section 2.3, labour productivity is expected to be an endogenous variable. To correct for simultaneous equation bias in the estimated coefficients because of the correlation between labour productivity and the error term, we have applied the method of Two-Stage Least Squares (2SLS). The 2SLS estimates are shown in column (4) of Table 4. In the first stage labour productivity is regressed on all the exogenous variables in the model. In the second stage sickness absence is regressed on the (other) exogenous variables and the predicted values of labour productivity obtained from the regression performed in the first stage.

To examine the selectivity bias in the estimates based on the KVZ-VZA-EWL-PS data introduced by linking the data with the PS, column (1) in Table 4 gives the estimates for the manufacturing companies based on the KVZ-VZA-EWL data. The difference between the estimates in the first and second columns of Table 4 gives an indication of the possible presence of selectivity bias. We find differences in significance of the considered explanatory variables for the following variables: percentage of unmarried workers, percentage of male workers, percentage of workers with regular working times, sick leave insurance and sector of economic activity. This implies that the estimated coefficients for these variables based on the KVZ-VZA-EWL-PS data are probably subject to selectivity bias, and therefore in this section we do not discuss the effect of the above-mentioned variables.

Table 4
OLS and 2SLS estimates of the logarithmic regression model for sickness absence (excluding maternity leave), manufacturing companies, 1996 [1)]

Variable	KVZ-VZA-EWL	KVZ-VZA-EWL-PS		
	(1) OLS	(2) OLS	(3) OLS	(4) 2SLS
Constant	3.052 (1.51)	–0.311 (0.11)	0.422 (0.15)	–0.068 (0.02)
Personal characteristics				
% unmarried	0.0027 (1.72)	0.052 (2.12)**	0.005 (1.98)**	0.005 (1.95)**
% male	–0.0017 (1.40)	–0.033 (2.29)**	–0.002 (1.66)	–0.003 (1.41)
% younger than 35 years	–0.0020 (1.38)	–0.017 (0.92)	–0.002 (0.93)	–0.001 (0.63)
Job characteristics				
Mean hourly wage (log)	–0.232 (2.89)**	–0.188 (1.88)*	–0.091 (0.95)	–0.187 (0.68)
Mean tenure (log)	0.0493 (1.59)	0.046 (0.13)	–0.005 (0.13)	0.005 (0.11)
% full-time workers	0.0023 (1.54)	0.006 (0.33)	0.0001 (0.07)	0.0004 (0.26)
% workers with regular hours	–0.0009 (1.46)	–0.015 (2.44)**	–0.002 (3.10)**	–0. 002 (3.06)**
Mean working hours (log)	–0.208 (0.75)	0.303 (0.76)	0.342 (0.91)	0.279 (0.74)
Company characteristics				
Size (log)	0.157 (11.05)**	0.092 (5.25)**	0.056 (2.18)**	0.069 (1.72)
Sick leave (re)insurance	–0.005 (0.15)	–0.066 (1.81)*	–0.076 (2.12)**	–0.061 (1.20)
Labour productivity (log)			–0.182 (4.88)**	0.077 (0.11)
Export share (log)			0.018 (1.49)	0.022 (1.42)
Market share (log)			0.051 (2.59)**	0.029 (0.50)
Capital intensity (log)			–0.055 (2.46)**	–0.113 (0.74)
Concentration ratio (C4, log)			–0.048 (1.70)	–0.041 (1.28)
Sector of economic activity (3-digit level): [2)]				
Food and stimulants	–0.202 (2.98)**	–0.042 (0.34)	0.110 (0.90)	0.027 (0.11)
Textiles and apparel	–0.128 (1.48)	–0.052 (0.38)	–0.081 (0.59)	–0.070 (0.50)
Leather	–0.515 (3.14)**	–0.496 (2.26)**	–0.585 (2.54)**	–0.581 (2.59)**
Wood	–0.151 (1.26)	0.077 (0.53)	0.022 (0.15)	0.038 (0.25)
Paper and publishing	–0.266 (3.78)**	–0.036 (0.29)	0.085 (0.70)	0.036 (0.21)
Petroleum	–0.417 (2.61)**	–0.262 (1.06)	0.070 (0.29)	–0.038 (0.11)
Chemicals	–0.382 (4.93)**	–0.250 (2.06)**	–0.061 (0.49)	–0.150 (0.59)
Rubber and plastic	–0.183 (2.21)**	0.012 (0.09)	0.087 (0.71)	0.053 (0.35)
Other non-metallic mineral	–0.157 (1.85)*	0.026 (0.21)	0.132 (1.02)	0.106 (0.75)
Metal	–0.177 (2.63)**	0.069 (0.06)	0.029 (0.25)	0.010 (0.08)
Machinery and equipment	–0.345 (4.92)**	–0.171 (1.41)	–0.176 (1.49)	–0.195 (1.55)
Computers, electrical apparatus	–0.340 (4.42)**	–0.210 (1.70)	–0.182 (1.46)	–0.204 (1.50)
Transport equipment	–0.097 (1.21)	0.041 (0.30)	0.009 (0.07)	0.015 (0.11)
R^2	0.162	0.119	0.155	0.121
Number of companies	1,743	1,281	1,281	1,281

[1)] Absolute values of t-values are mentioned in brackets.
** = significantly different from zero at the 95% level;
* = significantly different from zero at the 90% level.
[2)] The sector Culture or other services is excluded to avoid multicollinearity.

With respect to the effect of personal and job characteristics, the findings based on the data set KVZ-VZA-EWL-PS lead us to conclude that age has an insignificant influence on absence in the manufacturing sector as opposed to the private sector. The estimation results for the manufacturing sector in column (2) of Table 4 show a significant negative effect of mean hourly wage on the absence rate. However, after correction for the effect of some complementary company characteristics (such as labour productivity), the influence of the wage rate turns out to be insignificant (see column (3)). There appears to be no significant relationship between worker absence on the one hand and the percentage of full-time workers, the mean tenure and the mean working hours on the other hand for manufacturing companies.
When we examine the effect of company variables for the manufacturing sector, it follows from Table 4 that the addition of the company characteristics labour productivity, export share, market share, capital intensity and concentration ratio to the absence equation increases the R^2 by only 0.2% (compare the second and fourth columns). Thus, these company variables play a minor role in the determination of absence. From the OLS estimates (in the third column) we can conclude that company size, labour productivity, market share and capital intensity have a significant effect on worker absence. After correction with the 2SLS method for simultaneous equation bias in the estimated coefficients, we find that the estimated standard errors are higher and the influence of these company variables turns out to be insignificant (see column (4)). For the other included company variables – export share and concentration ratio – we also do not find a significant influence.
If we look at the influence of labour productivity, it can be seen from the OLS estimates that its effect on absence is significant and negative. The negative coefficient reflects the strong causal relationship from worker absence to labour productivity. After correction for simultaneous equation bias with 2SLS, the influence of labour productivity on absence turns out to be positive but insignificant.

5. Conclusions

Worker absence accounts for a lot of lost working time and therefore has implications for company productivity. In the Netherlands, 4.6% of total available working days in the fourth quarter of 1996 were lost as a result of absent employees. Research into the determinants of sickness absence is important to set out government policy aimed at reducing the number of days lost because of illness. With the aid of multiple regression analysis, this article has investigated the influence of personal, job and company characteristics on the differences in absence rates between companies. The analysis is based on linked cross-sectional data on individual companies in the Dutch private sector for the year 1996. The estimation results showed that company size, sector of economic activity, age of the workers, working schedules (regular/irregular/shift work), mean hourly wage and number of working hours have a significant influence on the number of days lost from work because of illness. For other personal, job and company characteristics (such as labour productivity) we do not find a significant effect on sick leave.
Finally, a limitation of the presented results should be pointed out. The regression estimates are based on cross-sectional data and may be biased because of possible correlation between the explanatory variables and the error term due to omitted variables. Only panel data of companies would enable us to deal more satisfactorily with the problem of unobservable fixed effects (such as health status, education level of workers, working conditions). Future research should concentrate on the analysis of linked panel data of individual companies.

For further information or comments on this contribution, please contact Martin Boon: mbon@cbs.nl.

References

Arnoldus, F., 1997, Electronic supply of data for labour statistics. In: *Netherlands Official Statistics* 12 (3), pp. 60–68. Statistics Netherlands, Voorburg.

Brown, S., F. Fakhfakh and J.G. Sessions, 1999, Absenteeism and Employee sharing: an empirical analysis on French panel data, 1981–1991. In: *Industrial and Labor Relations Review* 52 (2), pp. 234–251.

CBS, 1998, *Samenvattend overzicht van de industrie 1996-1997* (An overview of the manufacturing sector 1996–1997), Statistics Netherlands, Voorburg.

Coles, M.G. and J.G. Treble, 1996, Calculating the price of worker reliability. In: *Labour Economics* 3, pp. 169–188.

Fase, M.M.G. and L.M. Keijzer, 1991, Ziekteverzuim en conjunctuur (Sick leave and business cycle). In: *Economisch-Statistische Berichten* (10 March), pp. 372–375.

Groot, W., 1998, Oorzaken van langdurig ziekteverzuim van oudere werknemers (Causes of prolonged sickness absence of older employees). In: *Maandschrift Economie*, 62, pp. 309–318.

Keij-Deerenberg, I.M., 1996, Kwartaalonderzoek ziekteverzuim; methode en eerste uitkomsten (Quarterly sick leave survey: method and first results). In: *Sociaal-economische maandstatistiek* (December), pp. 20–23. Statistics Netherlands, Voorburg.

Kenyon, P. and P. Dawkins, 1989, A time series analysis of labour absence in Australia. In: *Review of Economics and Statistics* 71 (2), pp. 232–239.

NIA, 1996, *Berekening van ziekteverzuim: standaard voor verzuimregistratie* (Calculation of sick leave: a standard for absence registration), Nederlands Instituut voor Arbeidsomstandigheden, Amsterdam.

Otten, F. and H. Schmeets, 1995, Werkbelasting en ziekteverzuim (Work load and sick leave).In: *Economisch-Statistische Berichten* (19 April), pp. 383–385.

Smulders, P.G.W., 1984a, *Bedrijfskenmerken en ziekteverzuim in de jaren zestig en tachtig: een vergelijkende studie* (Company characteristics and sick leave in the sixties and eighties: a comparative study). Nederlands Instituut voor Praeventieve Gezondheidszorg TNO, Leiden.

Smulders, P.G.W., 1984b, *Balans van 30 jaar ziekte-verzuimonderzoek: de resultaten van 318 studies samengevat* (Overview of 30 years sick leave research: summary results of 318 studies). Nederlands Instituut voor Praeventieve Gezondheidszorg TNO, Leiden.

Toor, L. van, and B. Schaafsma-Harteveld, 1997, Werknemers-gegevens uit de verzekerdenadministratie: een eerste analyse (Employee data from the registration system of the social security funds: a first analysis). In: *Sociaal-economische maandstatistiek* (November), pp. 24–27. Statistics Netherlands, Voorburg.

Vistnes, J.P., 1997, Gender differences in days lost from work due to illness. In: *Industrial and Labor Relations Review*, Vol. 50 (2), pp. 304–321.

A brief overview of imputation methods applied at Statistics Netherlands

Ton de Waal

Abstract

One of the aims of the AutImp project, a European project under the fourth Framework, is the development of software for automatic imputation based on an innovative algorithm. In order to develop an imputation algorithm that can be applied in the day to day practice of various national statistical institutes, it is important to have some idea of imputation techniques that are currently applied at these institutes. The present article aims to give a brief overview of the imputation techniques that are currently applied at Statistics Netherlands.

Keywords: deductive (or logical) imputation, donor imputation, historical imputation, hot deck imputation, mean imputation, predictive mean imputation, ratio imputation, regression imputation, consistency, edit checks, AutImp

1. Introduction

AutImp is a project under the fourth Framework programme DOSIS. Partners in the project are the University of Southampton (UK), the Office for National Statistics (UK), Statistics Finland, Instituto Nacional de Estatística de Portugal and Statistics Netherlands. The AutImp project has two aims: evaluation of existing software for automatic imputation and the development of new software for this purpose based on an innovative algorithm. In order to develop an imputation algorithm that can be applied in the day to day practice of national statistical institutes, it is important to have some idea of imputation techniques that are currently applied at these institutes. The present article gives a brief overview of the most commonly used imputation techniques currently applied at Statistics Netherlands.

Statistics Netherlands is a large and complex organisation. It conducts many different kinds of surveys and uses many different techniques to process the resulting information. In particular, many different kinds of imputation techniques are applied to impute for missing values. As a complete description of all imputation techniques applied at Statistics Netherlands would require an intensive investigation lasting many months, we have opted for a more modest approach. The information contained in the present article has been obtained through Statistics Netherlands' intranet, from research reports, and through the author's own experience at the bureau. Although the study is thus limited in its extent, I hope it may serve as a starting point for Statistics Netherlands to pursue a more complete description of imputation techniques it applies.

The study distinguishes between economic and social surveys. Section 2 describes imputation methods applied for the former, and Section 3 for the latter surveys. Section 4 concludes with a few remarks about mass imputation. For an explanation of the terminology used in the paper we refer to Kalton and Kasprzyk (1986), Kovar and Whitridge (1995) and Schulte Nordholt (1998).

2. A brief overview of imputation methods applied for business surveys

2.1 Imputation

For enterprise data several kinds of imputation techniques are used at Statistics Netherlands. Below we describe the use of deductive imputation, regression imputation (including mean imputation, ratio imputation and historical imputation) and hot deck donor imputation. These techniques are applied to impute for both missing items and missing units.

Deductive, or logical, imputation is frequently applied at Statistics Netherlands, for instance, when one component of a total is missing. It is also applied when certain logical relations should hold true. For example, when the total of several non-negative items is reported to be zero, and the values of the items themselves are missing, the value zero is imputed for each of these items. Deductive imputation is also used to correct systematic errors. For example, financial data often have to be specified in thousands of guilders, but respondents report in guilders. In such a case the reported financial data are divided by 1,000.

Many of the imputation techniques applied at Statistics Netherlands belong to the class of regression imputation techniques. The regression models used are usually rather simple. In many cases a (weighted) class mean is imputed, in the *automation survey* for example. Mean value imputation is the simplest case of regression imputation, the regression model consists only of the constant term. The classes are not based on the data themselves, but on the output. That is, the publication cells of the tables that are released define the classes. In some cases the overall mean instead of the class mean is imputed, for instance when there are only a few observations per publication cell.

Another often applied technique is ratio imputation, again a very simple case of regression imputation. Here the regression model consists of a single predictor and the corresponding regression coefficient, and defines no constant term. At Statistics Netherlands the predictor used is mainly selected on the basis of subject-matter knowledge. In most cases data analysis is only used to justify the chosen predictor, not to select the best predictor. For each class, again defined by the publication cells, the regression coefficient is calculated. The predictor is the same for all classes. Like mean imputation, in some cases the regression coefficient is not calculated per class, but instead the overall regression coefficient is calculated. In many cases the number of employees of an enterprise is used as predictor, often taken from the *survey on employment and wages*.

Ratio imputation is, for instance, applied for the *order statistics of soil engineering, hydraulics engineering and road construction* to impute for missing units. For these statistics the value of the predictor is obtained from the same enterprise but from a different data source, e.g. from a VAT (value added tax) data set obtained from the tax department.

Historical imputation is also often applied at Statistics Netherlands. A prerequisite for applying this technique to a particular unit is that data on this unit from a previous period are available. In its basic form values reported by a unit from a previous period are used to impute for missing values in the current period. In a slightly more advanced form the values of a previous period are adjusted by a trend, usually based on other records in the same imputation class. This approach is for instance used to impute for missing values in the *annual production statistics of retail trade* (see Booleman, De Graaff and Verboon, 1995).

To impute for missing values in annual statistics, data from corresponding monthly or quarterly statistics are often used if available. This approach is used for instance for the *petroleum industry statistics*.

Imputation based on time series is also applied at Statistics Netherlands, but only rarely. The *foreign trade survey* uses moving averages to impute for some kinds of missing values. Some other kinds of missing values are imputed for by using exponential smoothing (see Diederen and Michels, 1996; Michels, 1996).

For the *construction industry survey* the so-called EDI-mix is used to impute for missing values caused by unit non-response (see Ravestijn). The term EDI-mix refers to the fact that internal data sources and external registers are used to guide the imputation process. The survey consists of about one hundred mostly numerical variables. For enterprises which did not respond, the value of either the turnover or the number of employees of that enterprise is determined by re-contacting the enterprise, or using data from another source. The 'structure' of the enterprise, i.e. the proportions between the various unknown variables and the observed variable, is estimated by using data from either the previous year or by using the average structure of the responding enterprises. The missing values are then imputed by preserving that structure. Below we illustrate the procedure used for the *construction industry survey* by means of a simple example.
Suppose the numbers of employees $N^{(t)}$ and $N^{(t-1)}$ of the enterprise in this period *t* and the previous period *t*-1, respectively, are observed. Suppose further that the value of variable *i* of the enterprise in the previous year is given by $x_i^{(t-1)}$. The imputed value of variable *i* of the enterprise in this year, $x_i^{(t)}$, is then calculated as

$$x_i^{(t)} = N^{(t)} \times \left(\frac{x_i^{(t-1)}}{N^{(t-1)}} \right) \tag{2.1}$$

When the proportions $x_i^{(t-1)} / N^{(t-1)}$ are unknown, the imputed value of variable *i* of the enterprise in this year is calculated as

$$x_i^{(t)} = N^{(t)} \times \left(\frac{\overline{x}_i^{(t)}}{\overline{N}^{(t)}} \right) \tag{2.2}$$

where $\overline{x}_i^{(t)}$ is an average value of variable *i*, and $\overline{N}^{(t)}$ an average number of employees. Both averages are taken over all responding enterprises in this year. In practice, the method applied for the *construction industry survey* is slightly more complicated, but the main idea is as described above.

Cold deck imputation, in the sense that data for an enterprise available in one data set are used to impute for data for the same enterprise in another data set, is often used. It is applied for instance in the *statistics of mechanisation in agriculture and horticulture*, where data from the so-called *agriculture census* are used to impute for missing values.
Statistics Netherlands rarely applies hot deck imputation methods for economic data. CherryPi, a computer program for automatic edit and imputation of numerical data (see below and De Waal, 1996), has been adapted in order to apply nearest-neighbour hot deck imputation to impute for missing data from the *environmental costs survey* (see Evers, 1998). Imputation classes can be specified; only records from the same imputation class as a record with missing data may be used as donor record for that record. The method has been tested extensively on data from the *environmental costs survey*. Data automatically edited and imputed by means of CherryPi have been compared with data edited and imputed in the traditional manner (see Evers, 1998), and the conclusion was that the quality of the CherryPi data is sufficiently high. Nevertheless, the adapted version of CherryPi with nearest-neighbour hot deck imputation is not applied in practice mainly because this version of CherryPi was developed especially for this particular survey and is therefore not general enough to be used for other surveys. Moreover, this particular version of CherryPi is not (yet) supported by the informatics department at Statistics Netherlands.
Apart from the experiments with the above-mentioned special version of CherryPi no other applications of hot deck imputation methods to numerical data from economic surveys at Statistics Netherlands are known to the author of the present paper.
Statistics Netherlands not only imputes for item non-response, but often also for unit non-response. Imputation of unit non-response is applied for practical reasons. Several different economic surveys are stored in a large database called the *Microlab*. To allow easy maintenance and easy use of the database, unit non-response is treated by imputation rather than by weighting.

2.2 Edit checks

Although in many cases edit checks are not taken into account when imputing for missing numerical values, in some cases they are. Israëls (1996) reports that for the so-called *summary of manufacturing* an imputation method is used that ensures that all edit checks are satisfied after imputation. Five edit checks are defined for this particular survey. All five are balance edits, i.e. edit checks stating that the sum of values of certain variables equals the sum of values of certain other variables. Imputation is carried out in two steps: first regression imputation is used to impute for missing values, and secondly the resulting record is made consistent by (slightly) modifying its values.

Suppose the (imputed) value of variable *j* of enterprise *i* after step 1 are given by $\hat{y}_{ij}$. We construct a record with values $\tilde{y}_{ij}$ that are close to $\hat{y}_{ij}$, and that satisfies all edit checks. To construct such a record the following objective function is minimised

$$\sum_j w_{ij} (\tilde{y}_{ij} - \hat{y}_{ij})^2 \tag{2.3}$$

subject to the constraint that the $\tilde{y}_{ij}$'s satisfy the five balance edits. Here the w_{ij}'s are fixed weights that measure the relative costs of modifying the values of the variables. All values may be modified during the second step, not only the imputed ones.
Some departments apply, or are planning to apply, CherryPi or SLICE (see John, 1997; Evers, 1998; Houbiers, 1999; Houbiers, Quere and De Waal, 1999; Schulte Nordholt and De Waal, 1999). SLICE is a general software package for editing and imputation that is currently being developed by Statistics Netherlands (De Waal and Wings, 1999). It will contain several edit and imputation modules. An example of such a module is an improved version of CherryPi.
CherryPi, or the CherryPi module of SLICE, can automatically identify implausible values in a record, set these to missing, automatically impute for missing values (the original missing values and the values that were set to missing because they were identified as being implausible) according to a regression model, and automatically adapt the imputed values in such a way that the resulting record is consistent with the specified edit checks. The edit checks that can be handled by CherryPi are linear equalities and linear inequalities.
To identify implausible values in a record automatically, CherryPi applies a generalised version of the Fellegi-Holt paradigm that states that a record should be made to satisfy all edit checks by changing the values of the fewest possible (weighted) number of variables (see Fellegi and Holt, 1976; De Waal, 1996; De Waal and Wings, 1999). To automatically impute for missing values the user can specify a general regression model. To modify the imputed values an objective function is minimised under the constraint that the resulting record satisfies all edit checks. Suppose that the imputed value of variable *j* of enterprise *i* after automatic imputation is given by $\hat{y}_{ij}$. Again we construct a record for record *i* with values $\tilde{y}_{ij}$ that are close to $\hat{y}_{ij}$, and that satisfies all edit checks. To construct such a record the following objective function is minimised

$$\sum_j w_{ij} \, | \tilde{y}_{ij} - \hat{y}_{ij} \, | \tag{2.4}$$

subject to the constraint that the $\tilde{y}_{ij}$'s satisfy the linear inequalities and equalities defined by the edit checks. Here the w_{ij} 's are again fixed weights measuring the relative costs of changing the values of the variables. Only the values that have been imputed may be modified during this step. Note that it is indeed possible to obtain a

consistent record in this way, i.e. by modifying only the imputed values. This is guaranteed, because the (generalised) Fellegi-Holt paradigm is used to identify the implausible values.

There are several differences between this approach and the approach by Israëls (1996) sketched above. Firstly, they use different objective functions; secondly the approach used in CherryPi allows a broader class of edit checks than the approach by Israëls; and thirdly, CherryPi only modifies the imputed values, whereas in the approach by Israëls non-imputed values may also be modified.
In the applications of CherryPi, for instance in the Dutch *labour costs survey* (Schulte Nordholt and De Waal, 1999) and the *survey on environmental costs* (Houbiers, Quere and De Waal, 1999), the regression models that were used to impute for missing values were quite simple. For most variables a mean value was imputed. For the other variables a simple regression model with only one predictor was used.

3. A brief overview of imputation methods applied in social surveys

3.1 Imputation

Schulte Nordholt (1998) describes the imputation strategy that was applied to impute for missing values of two surveys, the *housing demand survey* (see also Schulte Nordholt and Hooft van Huijsduijnen, 1997) and the *structure of earnings survey* (see also Schulte Nordholt, 1997), both conducted by Statistics Netherlands.
For the *housing demand survey* the variables were divided into groups, such as 'household', 'current dwelling', 'previous dwelling', 'respondent's socio-economic position', etc. Imputation was carried out per group of variables. Within a group of variables, those with the lowest percentages of missing values were imputed first. Discrete variables were mainly imputed by means of the random hot deck method, continuous variables by means of the random hot deck method or by means of predictive mean matching (see Schulte Nordholt, 1998). Related variables in different groups were imputed by record matching, or the common donor rule as this technique is also called (see Schulte Nordholt, 1998). So, only one donor record was used to impute for all missing values on related variables per record with item non-response. Variables already imputed were sometimes used as covariates for imputation of missing values not yet imputed. An important reason for this approach was to ensure internal consistency of the resulting imputed record. Schulte Nordholt (1998) gives an example in which the age of both the respondent and his/her partner are missing. In such a case the imputed value of the first variable was used as covariate during imputation of the second variable in order to prevent very unlikely age combinations.
The Dutch *structure of earnings survey* is created by matching three data sources: the *survey on employment and wages*, the registration system of the social security funds, and the *labour force survey*. A subset of the variables available from the three sources is selected for the *structure of earnings survey,* and these are used in the matching, imputation and weighting processes. Only exact matchings between the three data sources are used. After matching the sources the problem of missing values in the resultant data set arises. For some variables this problem is solved by imputation, for others by weighting. To impute for missing values the sequential hot deck method was applied per imputation class. Related variables were imputed simultaneously using the same imputation model in order to avoid the introduction of inconsistencies within an imputed record. The sequential not the random hot deck method was applied, because the number of records was considered too large to use the latter method efficiently. A random component was introduced in the imputation process by putting the potential donor records in a random order before the actual imputation took place.
Missing data from the *structure of earnings survey* were also imputed by means of a neural network (see Heerschap and Van der Graaf, 1999). In particular, missing values of the variable 'gross annual wage' were imputed by means of a feed forward multi-layer perceptron neural network. The best neural network that has been tested resulted in imputations of reasonable quality. However, at the same time it is doubted whether the quality of the resulting data set is sufficiently high to base publication figures on it. Even the best neural network needs to be refined in order to be applied in practice. More research is required before neural network imputation will be used at Statistics Netherlands.
Schulte Nordholt (1998) also describes the imputation strategy that was applied to impute for missing values of the *European Community Household Panel Survey* (ECHP) (see also Schulte Nordholt, 1996), which includes the Dutch *Socio-economic panel*. The ECHP focuses on income and the labour market, but includes topics such as health, education, housing and migration. The survey is carried out at the household level. The advantage of the ECHP over cross-sectional surveys is that it allows the study of income mobility patterns. However, the fact that it is a panel survey complicates the imputation process. Imputing each wave separately from the previous waves might lead to strange income shifts. Instead of this naive imputation strategy a better strategy would be to impute all available waves simultaneously. However, as this imputation strategy would lead to changes in the results of all previous waves after the imputation of a new wave, a simpler strategy was adopted: imputations of previous waves are taken into account while imputing a new wave, but imputations carried out during previous surveys will not be adapted.
To impute the first wave of the ECHP the random hot deck imputation within classes was used, with the aid of the software package Surfox (see Hooft van Huijsduijnen en Van Zijl, 1989). Because income variables are very important in the ECHP, the imputation concentrates on these. In addition to variables such as sex and year of birth, several important labour variables, such as employment status and present occupation, were used as auxiliaries to impute for missing values of income variables. The method takes great care to avoid inconsistencies between related variables, such as gross and net income from labour per month. Record matching was used to impute gross and net income from labour per month. Gross income was used as the main variable for imputation and the hot deck within classes method was used to impute for its missing values. If the value of the related variable net income from labour per month was also missing in a record, this was imputed from the same donor record to avoid inconsistencies.
Hilbink et al. (1995) report a case in which random hot deck imputation is applied to the *annual survey on employment and wages.* For practical reasons, in particular to speed-up the imputation process, not all potential donor records were used. For each imputation class a limited number of randomly selected donor records were stored in a so-called donor table. When missing values in a recipient record were to be imputed, a donor record from the donor table corresponding to the imputation class of the recipient record was chosen randomly. If no donor record was available to impute for a missing value, a value was selected according to the observed probability distribution of that variable.
For continuous variables in social surveys Statistics Netherlands often uses linear regression imputation. It hardly ever applies non-linear regression imputation. Several regression imputation models were considered for the *public libraries survey*, namely ratio imputation, linear regression imputation, and non-linear regression imputation (cf. Israëls and Pannekoek, 1999). A rather unusual - at least for Statistics Netherlands - kind of linear regression model was examined for this particular survey, namely

$$y = \beta x + \alpha x^2 + \varepsilon \tag{3.1}$$

where y is the dependent variable, x is the independent variable, and ε, is a stochastic error term with expectation zero and variance. σ^2
The following non-linear regression model was investigated:

$$y = \beta x^{\alpha} + \varepsilon \tag{3.2}$$

where *y*, *x*, and ε are the same as above. For this particular data set the linear imputation models appear to give better results than the non-linear model.

3.2 Edit checks

As we saw in Section 2, for economic data edit checks have to be taken into account explicitly during the imputation phase to guarantee that the imputed records satisfy these edit checks. In most cases, edit checks do not have to be taken into account explicitly for social data. Whereas for economic data consistent imputation, i.e. imputation in such a way that the resulting records satisfy the edit checks, is achieved in two steps (an imputation step using an imputation model, and a modification step) at Statistics Netherlands, for social data it is generally achieved in only one step. Two important reasons for this difference between economic and social data are:

– Statistics Netherlands collects a substantial part of its social survey data by means of *Blaise,* an integrated survey processing system developed by the bureau for computer-assisted interviewing. Inconsistencies in the data are detected during interviewing, and can be corrected immediately. Data collected by means of *Blaise* contain hardly any missing values or inconsistencies between the observed values. The latter fact implies that values are almost never set to missing during the editing phase because of such inconsistencies. For more information about *Blaise* we refer to the *Blaise Reference Manual* (1998) and the *Blaise Developer's Guide* (1998).
– The variables measured in social surveys are usually categorical ones. At Statistics Netherlands missing categorical values are mostly imputed by means of hot deck imputation. This method preserves certain multivariate relations between various variables and by controlling these multivariate relations, we can ensure that the imputed records satisfy all edit checks.

Statistics Netherlands has tested LINCE, a computer programme for automatic edit and imputation of categorical data (see Verboon, 1994), but decided against using it. We are now planning to adapt SLICE in such a way that it can automatically edit and impute mixed data, i.e. a mix of categorical and numerical data.

4. Mass imputation

Mass imputation has been discussed extensively at Statistics Netherlands. There has been no general census in the Netherlands for several decades, nor are there plans to conduct one in the near future. To obtain census-like data Statistics Netherlands plans to combine several surveys and administrative registers. One suggestion has been to mass impute all missing values, i.e. a complete record will be constructed for each inhabitant of the Netherlands, even though we often only have administrative data available for this person. Similarly, Statistics Netherlands plans to fill the so-called *Microlab* with data for all enterprises. Missing values, due to item or unit non-response, may be mass imputed.
A practical advantage of mass imputation is that weighting is no longer required to generate publication figures. A methodological advantage may be that auxiliary information is used more efficiently than when weighting is used naively. However, the statistical methods department has raised several methodological objections to this kind of mass imputation (see e.g. Kooiman, 1998), and has developed an alternative approach based on weighting (see Renssen and Kroese, 1999). At the moment of writing, it is not yet know which of the two approaches, mass imputation or weighting, will be adopted in practice.

For further information or comments on this contribution, please contact Ton de Waal: twal@cbs.nl.

References

Statistics Netherlands, 1998. *Blaise reference manual.* Heerlen.

Statistics Netherlands, 1998. *Blaise developer's guide*. Heerlen.

Booleman, M., C. De Graaff and P. Verboon, 1995. *Een beschrijving van het surveyproces bij de PS detailhandel* (A description of the survey process of the production statistics of retail trade). Internal report, Statistics Netherlands, Voorburg.

De Waal, T., 1996. *CherryPi: A computer program for automatic edit and imputation.* Internal report, Statistics Netherlands, Voorburg.

De Waal, T. and H. Wings, 1999. *From CherryPi to SLICE*. Internal report, Statistics Netherlands, Voorburg.

Diederen, B., and P. Michels, 1996. *Het gebruik van tijdreeksen bij het controle-, correctie- en publikatieproces na het herontwerp van de statistieken van de internationale handel* (The use of time series in the edit, correction and publication process after the redesign of the foreign trade surveys), Internal report, Statistics Netherlands, Heerlen.

Evers, I., 1998. *Nearest-neighbour hot deck imputatie in CherryPi* (Nearest-Neighbour Hot Deck Imputation in CherryPi).Internal report, Statistics Netherlands, Voorburg.

Fellegi, I.P. and D. Holt, 1976, A systematic approach to automatic edit and imputation. In: *Journal of the American Statistical Association*, 71, 17–35.

Heerschap, N. and A. Van der Graaf, 1999. *Een imputatietest met een neuraal netwerk op basis van de gegevens van het loonstructuuronderzoek* (A test of Imputation by means of a neural network on data of the structure of earnings survey). Internal report, Statistics Netherlands, Voorburg.

Hilbink, K., E. Hoogteijling, F. Van de Pol, and E. Schulte Nordholt, 1995. *Algoritme voor imputatie bij het JWL* (Algorithm for imputation in the annual survey on employment and wages). Internal report, Statistics Netherlands, Voorburg.

Hooft van Huijsduijnen, J. and A. Van Zijl, 1989. *Surfox (release 1.0), User's Manual* (in Dutch).

Houbiers, M., 1999. *Automatisch gaafmaken van de statistiek bedrijfsafvalstoffen 1996 met behulp van CherryPi* (Automatic editing of the industrial waste survey 1996 by means of CherryPi). Internal report, Statistics Netherlands, Voorburg.

Houbiers, M., R. Quere and T. De Waal, 1999. *Automatically editing the 1997 survey on environmental costs* (in preparation). Internal report, Statistics Netherlands, Voorburg.

Israëls, A., 1996. *Simultaneous imputation under balancing constraints.* Report, Statistics Netherlands, Voorburg.

Israëls, A. and J. Pannekoek, 1999. *Imputatie bij de statistiek van de openbare bibliotheken* (Imputation in the survey of public libraries). Internal report, Statistics Netherlands.

John, P., 1997. *Een eerste test van automatisch gaafmaken met CherryPi* (A first test of automatic editing by means of CherryPi). Internal report, Statistics Netherlands, Voorburg.

Kalton, G. and D. Kasprzyk, 1986. The treatment of missing survey data. In: *Survey Methodology*, 12, 1–16.

Kooiman, P., 1998. *Massa-imputatie: waarom niet?* (Mass imputation: why not?). Report, Statistics Netherlands, Voorburg.

Kovar, J.G. and P.J. Whitridge, 1995. Imputation of business survey data. In: *Business Survey Methods*. John Wiley & Sons.

Michels, P., 1996. *Voorspellingen met effeningsmethoden voor controle/correctie bij de internationale handel* (Predictions using smoothing methods for edit/correction of the foreign trade survey). Internal report, Statistics Netherlands, Heerlen.

Ravestijn, R., *Imputeren met behulp van EDI-mix bij de productiestatistiek bouwnijverheid* (Imputation by means of EDI-mix in the production statistics of the construction industry). Internal report, Statistics Netherlands, Voorburg.

Renssen, R. and B. Kroese, 1999. *Een methodologisch model voor het vullen van de reference database* (A methodological model for filling the reference database). Report, Statistics Netherlands, Heerlen.

Schulte Nordholt, E., 1996. *The used techniques for the imputation of wave 1 data of the ECHP*. Research report doc. PAN 66/96, Eurostat, Luxembourg.

Schulte Nordholt, E., 1997. *Imputation in the new Dutch structure of earnings survey (SES)*. Report, Statistics Netherlands, Voorburg.

Schulte Nordholt, E., 1998. Imputation: methods, simulation experiments and practical examples. In: *International Statistical Review*, 66, 157–180.

Schulte Nordholt, E. and J. Hooft van Huijsduijnen, 1997. The treatment of item non-response during the editing of survey results. In: *New Techniques and Technologies for Statistics II. Proceedings of the Second Bonn Seminar*, IOS Press, Amsterdam.

Schulte Nordholt, E. and T. De Waal, 1999. *Automatic editing in the Dutch labour costs survey using CherryPi.* Report, Statistics Netherlands, Voorburg.

Verboon, P., 1994. *Een beschrijving van het edit- en imputatiepakket LINCE; Versie 1.6* (A description of the edit and imputation program LINCE; Version 1.6). Internal report, Statistics Netherlands, Voorburg.

Comparing national housing situations within Europe

Jeanne Roijen

Abstract

Eurostat's European Community Household Panel Survey (ECHP), which is conducted among members of private households, covers a wide range of demographic, social and economic issues and makes it possible to give a description of the living conditions of residents (both individuals and households). One aspect of the living conditions is the housing situation. The present contribution describes the housing situation of people aged sixteen years and older, and their opinions about this situation in thirteen countries of the European Union.

Key words: housing, ECHP, living conditions

1. Introduction

Not many surveys offer the opportunity to compare the housing situation in the Netherlands with that in other countries. One survey that does offer such an opportunity is Eurostat's *European Community Household Panel Survey* (ECHP), a longitudinal survey that, in principle, is conducted yearly in the countries of the European Union. The survey is intended as a multi-dimensional monitor and research instrument for the European Union. It focuses on the improvement of the working conditions, living conditions, citizenship and social cohesion.

The ECHP is conducted by the participating countries themselves with the aid of a questionnaire drawn up by Eurostat, the statistical bureau of the European Union. This questionnaire was developed in close co-operation with the countries involved and can be adjusted to national research practices. For each country, the fieldwork, checks and weighting are carried out by national research institutes, usually the national statistical office. The national databases are combined by Eurostat to form one central database.

In the Netherlands, no special survey is held for the ECHP; the ECHP questions are integrated in the existing *Socio-economic Panel Survey* (SEP) which is conducted yearly by Statistics Netherlands among some five thousand households.

The first wave of the ECHP was held in 1994. The second wave, whose results are described in the present article, was conducted in 1995 among a total of some sixty thousand households comprising 130,000 people aged sixteen years or older. Thirteen countries of the European Union participated in this wave: Austria, Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain and the United Kingdom.

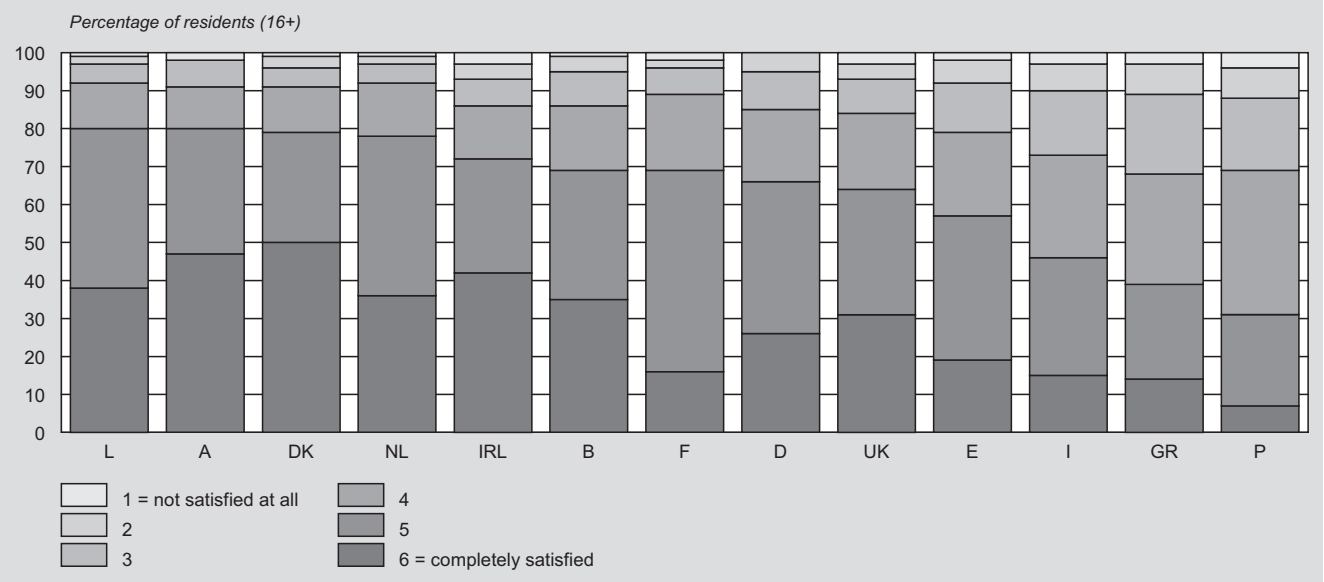
The figures presented in the present contribution are relative. Absolute figures – in many cases even integral figures – for the various countries are available from national sources and can be found in the United Nations' *Annual Bulletin on Housing and Building Statistics for Europe and North America* (See Table 2).

The following sections illustrate certain aspects of the housing situation of residents aged 16 and older across the thirteen countries participating in the survey.

2. Satisfaction with the housing situation

As Figure 1 shows, more than 75 percent of the residents in question in the Netherlands, Luxembourg, Austria and Denmark are satisfied, or even very satisfied (categories 5 and 6 in the figure), with their own housing situation. Fewer southern Europeans fall in these two categories: Spain (57 percent), Italy (46 percent), Greece (39 percent) and Portugal (31 percent). The other countries take a middle position, with some 60 to 70 percent of respondents reporting satisfaction and complete satisfaction with their own housing situation.
The high scores on the satisfaction item give the impression that the housing situation in the Netherlands (as well as in several other countries) is quite favourable. Below we shall see whether this is confirmed by other aspects of the housing situation.

Figure 1. Satisfaction with the housing situation, 1995



3. Dwelling type

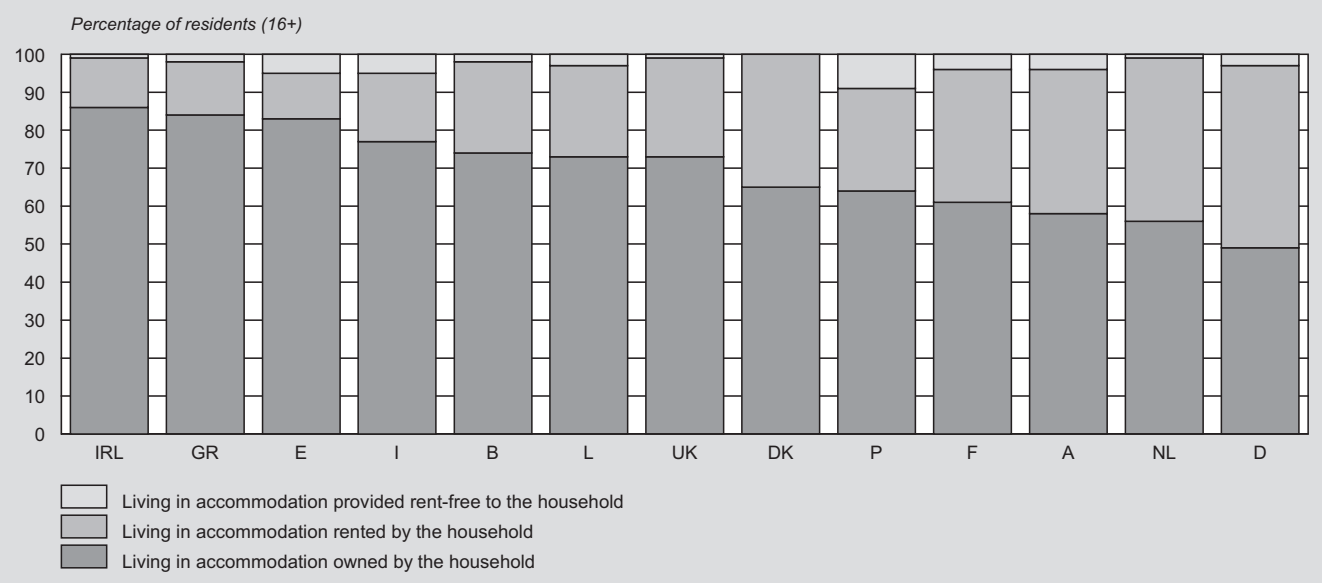
In the Netherlands, with its high population density and shortage of space, a detached single family house is regarded as a luxury, the dream home in fact for many people. The fact that so many Dutch people are so satisfied with their housing situation does not mean, however, that they live in such houses. As Figure 2 shows, the Netherlands ranks below all the other countries with regard to this type of home. Only sixteen percent of the population live in a detached house. The highest percentages in this respect can be found in Denmark (56), Portugal (49), Ireland (48) and Austria (47).
Most houses in the Netherlands are semi-detached or terraced houses; over half of the Dutch live in one of these two types. Only in the United Kingdom is this proportion higher (63 percent). In Ireland people live predominantly in either a terraced, semi-detached or a detached house (both 48 percent), there are hardly any flats and apartment blocks. These are quite common in other countries: Spain (62 percent of population in private households), Italy (61 percent) and Germany (52 percent). In the Netherlands the share of residents living in an apartment is rather small: four percent live in apartment blocks with fewer than ten apartments, and seventeen percent in larger apartment blocks.

4. Ownership and tenancy

In general, people looking to settle in a certain neighbourhood have the widest choice if they can buy a house. Surprisingly, the figures (Figure 3) contradict this reasoning. Even though a larger share of Dutch and German people reported being satisfied or very satisfied with their housing situation, a relatively small proportion of the population in these countries live in an owner-occupied house. While more than 80 percent of Irish people, Greeks and Spaniards live in a house owned by the household, in the Netherlands this is true only for 56 percent of the residents and in Germany for 49 percent. In the other countries the proportions range from 60 to 80 percent.

Figure 2. Type of dwelling, 1995

Figure 3. Home ownership, 1995



5. Household size

More than half of Dutch people live either alone or with only one other person. Only Denmark and Germany have similar proportions of one or two-person households. In the remaining countries over half the population live in larger households.
In each of the participating countries, more than one third of the people live in households consisting of three or four persons. In Greece, Italy and Portugal this is even true for more than half of the population. In Ireland and Spain relatively many people are part of a household consisting of five or more persons (respectively 33 and 27 percent). Such large households are not very common in the Netherlands, Denmark and Germany, accounting for no more than ten percent of the population.
In terms of the number of rooms (exc. kitchen) in the dwelling, Dutch residents live in larger houses than people in other countries. Whether these houses are also larger in terms of floor space is not known as no questions on this were included in the survey.

In all the countries the proportion of people living in a one-room dwelling is very small (no more than three percent of the population). Two-room dwellings are most common in Greece (20 percent), while in other countries they account for no more than 15 percent).

In the Netherlands more than 70 percent of residents aged 16 or over live in a dwelling with five rooms or more. This proportion is smaller in the other countries (Ireland 62 percent, Luxembourg 53 percent, other countries no more than 50 percent).

6. Comfort and deficiencies

How comfortable a house is to live in depends on many factors. In the survey several basic amenities in or outside the house were taken into account.

Figure 4. Household size, 1995

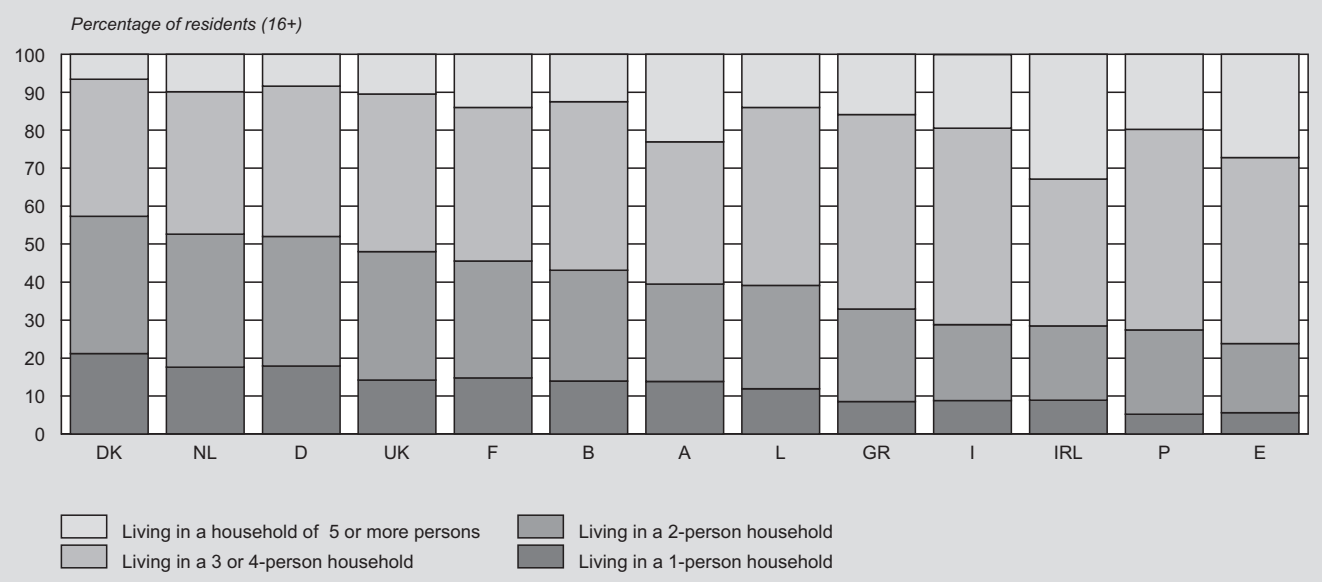
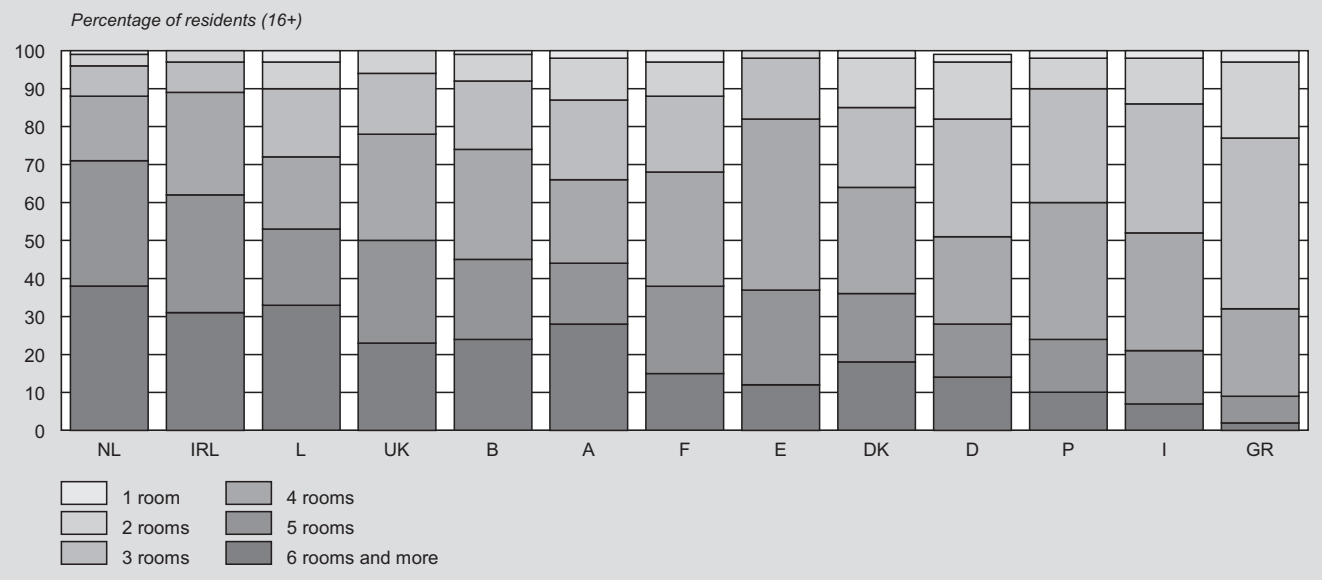


Figure 5. Size of dwelling (number of rooms, kitchen excluded), 1995



In most countries houses are equipped with hot running water, an indoor flushing toilet and a bath or shower, but:
– in Greece 69 percent and in Portugal 19 percent of the residents do not have hot running water at their disposal; the remaining countries all have percentages of five or lower in this respect;.
– in Portugal 12 percent of people do not have an indoor flushing toilet; in Greece this is six percent and in Austria five. In the other countries the share is no more than three percent;
– in Portugal 13 percent of the population in question do not have a bath or shower in the house, the same is true for four percent of Greeks. In the remaining countries the share is three percent or lower.

As we can see in Figure 6 two phenomena appear to be typically Dutch .
– The most noticeable is that in the Netherlands one third of the people live in a dwelling without a separate kitchen, whereas in the other countries the share is 11 percent at the most. It is hard to judge whether this can be labelled as the lack of an amenity. In the Netherlands it is quite usual to build houses, even the more expensive ones, with the kitchen integrated in the living room (a so-called 'open kitchen').
– In the Netherlands almost all dwellings are equipped with a place to sit outside (a garden or balcony, etc.); only three percent of the Dutch population mentioned the lack of such a facility. In other countries it is more common for people not to have a place to sit outside. In France, Spain, Portugal and Austria more than 20 percent of respondents reported the lack of such a facility.

The ECHP survey asked people whether they had problems with certain specified problems with their dwelling, such as shortage of space, not enough daylight, inadequate heating, a leaky roof, damp walls or rot in the floors or window frames. (See Table 1*).*

Dutch people turn out to have fewer problems with their houses than people in many other countries. More specifically:
– Relatively fewer people in the Netherlands do not have enough room at home (nine percent compared with 12 to 32 percent in

Figure 6. Lack of amenities, 1995

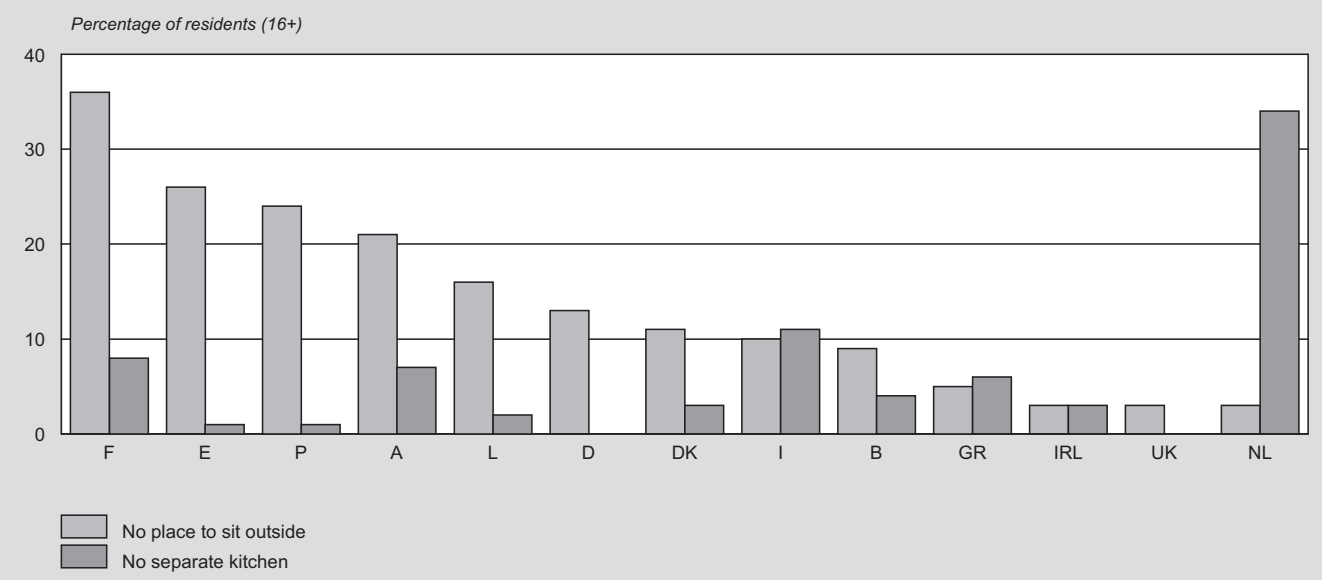


Table 1
Residents (16+) in private households by problems with the dwelling and by country, 1995

	Shortage of space	Too dark / not light enough	Inadequate heating facilities	Leaky roof	Damp walls, floors or foundation	Rot in window frames or floors
	% of residents					
Portugal	32	19	40	16	34	29
Greece	30	9	34	16	16	9
Spain	26	17	1	10	19	7
United Kingdom	22	8	10	4	15	14
Italy	19	10	16	6	5	6
Austria	18	6	7	4	11	6
Denmark	16	3	3	4	6	5
Ireland	14	3	7	4	9	7
France	14	9	10	6	17	10
Belgium	13	10	7	5	14	8
Germany	12	5	5	4	7	5
Luxembourg	12	4	6	6	9	5
Netherlands	9	5	6	5	12	10

Source: ECHP.

other countries). People in Portugal and Greece in particular feel they do not have enough space.
- People in Portugal (19 percent) and Spain (17 percent) think their dwellings are too dark. In the Netherlands this is a problem for only five percent of residents.
- In the south European countries, with exception of Spain, there is often a lack of adequate heating facilities: Italy (16 percent), Greece (34 percent) and Portugal (40 percent). In the Netherlands, Luxembourg, Germany, Denmark and Spain this is much less common (six percent at the most).
- Leaky roofs cause problems for households in Portugal, Greece and Spain, where 10 to 16 percent of people reported this complaint. In the other countries this is four to six percent.
- More often, damp is a problem, especially in Portugal; 34% of Portuguese residents complained of damp walls, floors or foundations. In the other countries this problem was less common (up to 19 percent). In the Netherlands it is 12 percent.
- The share of people reporting problems with rot in window frames and floors was higher in Portugal (29 percent) than in the other countries. In the United Kingdom it is 14 percent, in the Netherlands and France ten percent, and in the other countries the share is lower.

7. Costs for housing

The degree to which the housing costs are a burden on the household budget is also taken into consideration in the ECHP survey.
In Luxembourg, the Netherlands and Denmark by far most residents have no problems with the costs of their home; this is the case for 82, 71 and 69 percent respectively. In Spain and Italy only 14 percent and six percent respectively of residents do not have a problem with the housing costs. These are also the two countries with the most people reporting problems with housing costs; more than one third of the residents described the housing costs as a heavy burden on the household's budget.

Figure 7. Housing costs, 1995

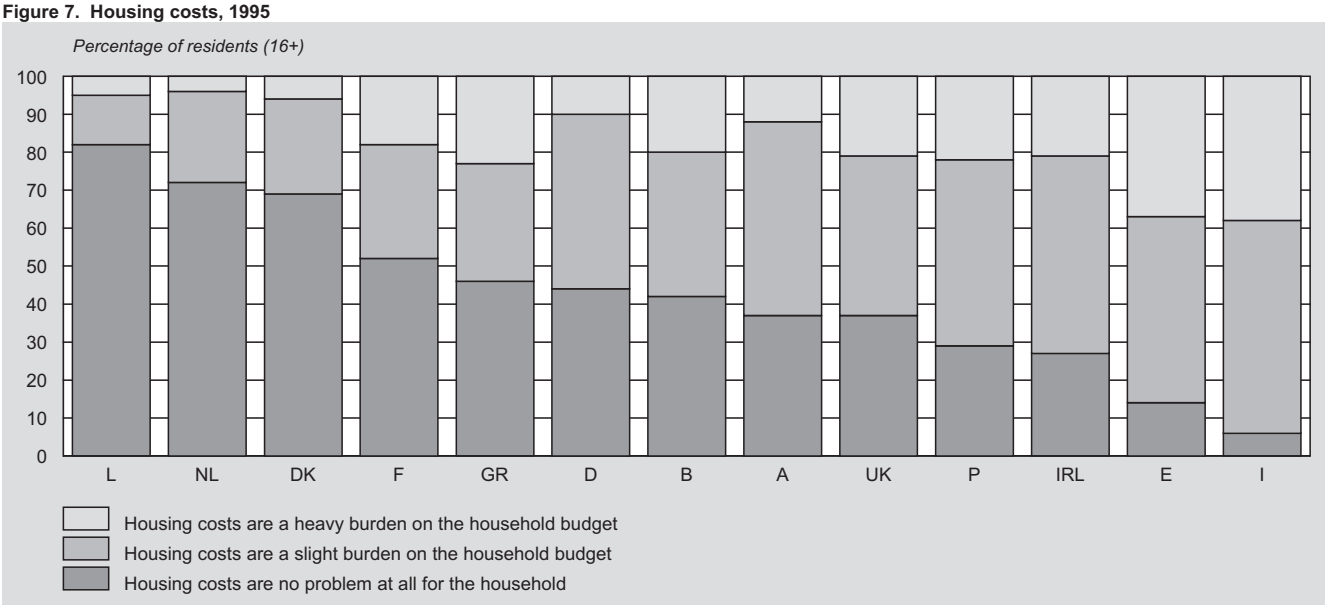
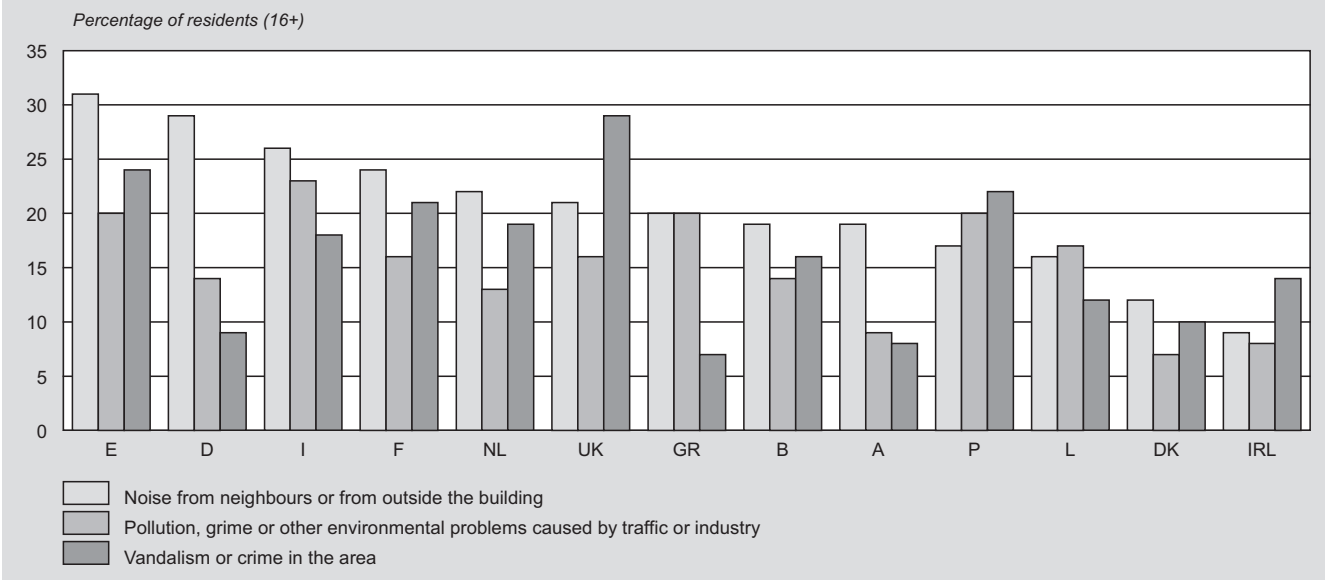


Figure 8. Problems with the neighbourhood, 1995



8. The neighbourhood

It is not only characteristics of the dwelling that makes people satisfied (or not) with their housing situation; the quality of the neighbourhood is also an important factor. The aspects of the neighbourhood included in this survey are:
– noise from neighbours or from outside the building;
– pollution, grime or other environmental problems caused by traffic or industry;
– vandalism or crime in the area.

As can be seen in Figure 8 people in all countries have to put up with these kinds of annoyance. Noise is mentioned most in Spain (31 percent) and least in Ireland (nine percent). In the Netherlands 22 percent of the residents have to put up with this kind of problem. Pollution, grime and other environmental problems caused by traffic or industry are less of an issue. In Italy 23 percent of respondents mentioned these, in Spain, Greece and Portugal 20 percent. In the other countries the share was between 7 and 17 percent.
In the United Kingdom 29 percent of the population experience vandalism or crime in the area. In Portugal, France and Spain this is 21 to 24 percent. The Netherlands ranks relatively high with respect to this problem (19 percent), and residents in Greece, Austria, Germany and Denmark are troubled least by such events (no more than ten percent) In the other countries the share is 12 to 18 percent.

9. Conclusion

Residents in the Netherlands are mostly satisfied or even completely satisfied with their housing situation. About half of Dutch people own their home, the other half rents accommodation. By far most of them live in a semi-detached or terraced house, relatively few in a detached house. Dutch homes are quite large in terms of number of rooms, and only few residents complain of a shortage of space. The houses almost always have a garden, patio or balcony. In general the Dutch have few problems with their homes, and when they do experience problems these are mostly damp walls, floors or foundations or rot in floors or window frames. Nearly a quarter of Dutch people have to put up with noise from neighbours or from outside, fewer people are annoyed by pollution and crime or vandalism in the neighbourhood. Most people in the Netherlands live in a household that has no problems with the housing costs.

The Netherlands is one of the four countries with the highest percentage of satisfied and very satisfied residents. Compared with people in the other surveyed countries, fewer Dutch residents live in a detached house and, apart from the United Kingdom, the Netherlands has the highest share of households in semi-detached and terraced houses. While Dutch households are smaller than those in all other countries except Denmark, they have more rooms at their disposal. Relatively few people in the Netherlands complain of a shortage of space and more people have an opportunity to sit outside (garden, balcony, etc.) than in the other countries.

A greater proportion of Dutch people live in dwellings that are light enough and have adequate heating facilities. As in other countries most people live in houses that have a sound roof . Problems with damp are mentioned more often than a leaky roof, but that goes for most of the countries. As far as rot in window frames, floors or foundations is concerned, the Netherlands is one of four countries that experience these problems most.

The Netherlands is one of the top five countries where neighbourhood noise is concerned. Vandalism or crime in the area are mentioned more often in the Netherlands than in many other countries. Compared with the other countries fewer Dutch people live in a house of their own. Together with Luxembourg and Denmark, in the Netherlands the housing costs are seldom a heavy burden on the household purse .

Finally, Table 2 presents basic data on housing in Europe, based on the United Nations' *Annual Bulletin of Housing and Building Statistics.*

For further information or comments on this contribution, please contact Jeanne Roijen: jren@cbs.nl.

Table 2
Basic data on housing in Europe

	Surface area in km²	Population	Households	Dwelling stock	Consumer price and rent indices	
					total	rent
	1997	1996	1996	1996	1996	1996
	x 1,000	*x mln*	*x 1,000*	*x 1,000*	*1990=100*	
Belgium	31	10	3,747 [5)]	3,974 [12)]	115 [16)]	124 [5)]
Denmark	43	5 [1)]	2,344 [6)]	2,447	111	120
Germany	357	82 [2)]	32,897 [7)]	35,954 [13)]	117 [17)]	138
Finland	338	5	2,390	2,391	116	130
France	544	58	23,286	28,221	113	124
Greece	132	10	3,198 [5)]	4,652 [5)]	204	266
Ireland	69	4	1,085 [8)]	1,107 [13)]	116 [18)]	108
Iceland	103	0 [3)]	93 [5)]	97 [6)]	124 [19)]	109
Italy	301	57	19,909 [5)]	25,028 [5)]	127 [5)]	142 [5)]
Luxembourg	3	0	145 [5)]	106 [5) 14)]	114 [5)]	135 [5)]
Netherlands	41	16	6,039 [8)]	6,366	117 [16)]	133
Norway	324	4 [1)]	1,752 [5)]	1,862	113	115
Austria	84	8	3,142	3,142 [12)]	118	135
Portugal	92	10 [4)]	3,113 [5)]	3,059 [5)]	131 [17)]	147
Spain	505	39	11,736 [5)]	17,160 [5)]	119 [20)]	131
United Kingdom	242	59	23,300 [9) 10)]	24,598	125	154
Sweden	450	9	3,794 [11)]	4,044 [7)]	116	143
Switzerland	41	7	2,813 [11)]	3,434	115	127

Sources: UN/ECE, Annual Bulletin of Housing and Building Statistics for Europe and North America, 1998.
Eurostat, Yearbook 1997.

* Consumer price totals exclude shelter (that means primarily without rent).
** Some of the rent indices relate to rent both in old and new dwellings, others are limited to rents in old dwellings only, others refer to rents including rates and water charges or even fuel and light and charges for repair and maintena i.e. sometimes the rent indices comprise what is generally referred to as housing costs.

1) Population on 1 January.
2) Population on 31 December.
3) Population on 1 December.
4) Provisional data.
5) 1991.
6) Population on 1 January 1997.
7) 1993.
8) 1994.
9) Great Britain, 1995/1996.
10) 1995/96.
11) 1990.
12) Data refer to the main residences.
13) 1995.
14) Occupied dwellings.
15) Including rent.
16) 1991 = 100.
17) Excluding housing costs.
18) Reykjavik.
19) 1992 = 100.

Dissertation on Consumer Price Index construction

On 28 January 2000, Jan de Haan, researcher at the Department of Consumer Prices was awarded a PhD by Rotterdam's Erasmus University for his dissertation *Empirical studies on consumer price index construction*.

The thesis distinguishes three steps in the calculation of a Consumer Price Index (CPI).

1. In the *first step* the target of the CPI has to be formulated, e.g. measuring inflation or, alternatively, a constant utility price (or cost of living) index, which measures the change in the minimum costs for a representative consumer (or household) needed to retain the same standard of living or utility.
2. Since inflation is only a general notion and utility cannot be measured directly, a certain index formula must be chosen as a *second step*. Such a formula is a mathematical expression where the prices and quantities of all commodities come into play. Index number theory gives some guidance as to the appropriate choice. Superlative indexes provide second-order approximations to the (true) cost of living index. Their most important feature is that they take into account the substitution between goods and services resulting from relative price changes. These index formulas make use of quantity or expenditure data relating to both the base period 0 and the current period *t*. In practice it takes some time before expenditure data are available. Timeliness of the CPI, which is generally viewed as an important aspect of its quality, has led most national statistical offices to adopt the Laspeyres price index as their target index or object of estimation. After aggregating all commodities into a number of commodity groups, the Laspeyres price index can be written as a weighted average of the various commodity group indexes, where the weights reflect base period expenditure shares.
3. In the *third step*, sampling procedures are needed to estimate the population value of the Laspeyres index. Ideally, the mean square error of the estimator should be minimised.

The three steps can be seen as aspects of a broader concept that might be used to assess the quality of official statistics. From the literature five quality components are taken: *relevance, accuracy, timeliness, coherence* and *availability*. Relevance concerns the choice of the statistical targets, e.g. units and population, variables and study domains. The first two steps address this quality component as well as the component of timeliness – taking the Laspeyres index formula as the primary CPI target. Accuracy refers to the need for estimating CPIs close to the true population value, taking cost constraints into account, and presenting measures such as standard errors or confidence intervals and describing possible sources of inaccuracy (including non-sampling errors). Coherence relates to sets of statistics, and takes into account how well these sets can be used together. Two sub-aspects are of special importance. If there is a single time series, like the CPI, there is comparability over time. If the statistics set comprises statistics for different domains with similar target characteristics, such as CPIs for different groups of households or for different countries, we have comparability over space. Lastly, availability refers to the way in which the statistics are disseminated and presented, and to the users' possibilities to acquire documentation on the published statistics.

The eight chapters of the thesis are divided into three parts. Part I deals with estimating CPI sub-index weights. Like most national agencies Statistics Netherlands obtains the weights using straightforward design-based estimators. There are some problems because of the relatively small sample size of the expenditure survey, which leads to large sampling errors, and the possibility of bias caused by underreporting in the survey. Part II deals with the use of bar-code scanning data in empirical index number research. Such scanner data offer challenging perspectives in this field, for two reasons. Firstly, all sorts of price index numbers, including superlative indexes, can be compiled using highly disaggregated data at the individual outlet level. Secondly, since the scanner data sets cover almost all products sold within specified categories, the 'true' population values of the price indexes can be calculated and confronted with index numbers based on samples. Part III is termed miscellaneous and contains work that does not fit into the first two parts.

Chapter 2 focuses on the estimation of the CPI's weights. One of the estimators considered uses a regression model applied to expenditure survey micro-data. In the standard model-assisted regression approach, the auxiliary variables are assumed to be known for every element in the population. Since in the case presented this information is lacking only survey data are used and the regression is run on sample data from outside the population in question as well. Although this model-based approach slightly lowers the standard errors of the CPIs, the introduction of bias (which cannot be measured directly) prevents its practical implementation.

Chapter 3 appreciates that national accounts consumption data are often superior to survey data. To provide users of official statistics with a coherent picture of expenditures and price changes, it is worthwhile to link CPI-weights for domains to the national accounts figures. This is done by allocating the differences between the national accounts figures and the survey estimates for the entire population on each expenditure category to the domains. It is argued that although an allocation method that minimises the variance of the resulting new domain expenditure estimates is to be preferred in theory, proportionally allocating the differences according to the initial survey estimates has some important practical advantages.

Chapter 4 addresses several issues in compiling the consumer price sub-index of coffee at the lowest aggregation level. It reports on a Statistics Netherlands pilot study which makes use of scanner data, and can be regarded as an introduction to the next two chapters. Preference is given to using the unit value index at some low aggregation level. This not only holds for the aggregation in space (i.e. over 'homogeneous' commodities), but also for the aggregation in time to calculate monthly average transaction prices. It is shown that, for coffee at least, the price concept is far more important than the choice of the index formula at the elementary level.

Chapter 5 discusses the sampling of commodities or items. To date, this issue has not attracted the attention it deserves, mainly because of the lack of data. Statistics Netherlands is currently using a judgmental selection method, which can be conceived of as a kind of cut-off sampling. The only way of knowing the bias of the item group indexes caused by this sampling method is by looking at the entire population of items belonging to the group. This is done using scanner data on three item groups. The cut-off selection method is compared with simple random sampling, stratified sampling and (systematic) sampling proportional to expenditure. Because standard variance estimation methods might be inaccurate, and also in order to calculate the bias that may exist even under random sampling, Monte Carlo simulations have been run to describe the sampling distributions. Surprisingly perhaps, cut-off sampling performs best in terms of the mean square error for coffee and toilet paper, whereas for nappies this method yields results very similar to sampling proportional to expenditure. As far as the author is aware this is the first study to supply empirical evidence in support of the use of cut-off commodity selection methods.

Chapter 6 addresses the estimation of price changes at the lowest, or elementary, aggregation level where quantity and expenditure data is lacking. In particular, it addresses the bias of three different price index estimators with respect to some ideal population index under various sampling designs. The bias is decomposed into substitution bias, target bias and small sample bias. The second component – often neglected in studies concerning price indexes at the elementary aggregation level – shows by how much the expectation of the estimator differs from the fixed-base operational target. The scanner data sets of Chapters 4 and 5 are used to quantify the bias for six items.

The CPI measures the change in market prices, thus including the effect of changes in indirect taxes like Value Added Tax and excise duties. Moreover, the Dutch CPI includes so-called consumption related taxes (such as motor vehicle tax) as separate expenditure categories. For various reasons users may wish to exclude the effect of taxes. Chapter 7 elaborates on the principles behind the computation of the Dutch net price index. A distinction is made between direct tax effects due to changes in taxes levied on consumer goods and services and indirect effects resulting from changes in taxes levied on intermediate products used in the production processes of enterprises. The assumptions underlying the net price index are discussed at length.

Chapter 8 goes into the construction of price indexes for insurance that provides some reimbursement on the repair or replacement of damaged or stolen goods. Two well-known methods are compared: a method based on direct observation of premiums and a method based on changes in repair and replacement costs and the costs associated with the services provided by insurance companies. The concepts underlying both methods are discussed and a number of difficulties that arise in practice are outlined. It is argued that the cost-based approach should not be used and that investment earnings on premiums should be excluded from the CPI.

For further information, please contact Jan de Haan: jhhn@cbs.nl.

International papers and contributions by Statistics Netherlands staff in 1999

Policy staff

Abrahamse, A.P.J. and J.R. Nobel. The Social Question: Statistics and Public Policy. In: F.G. Pyatt and M. Ward (eds.). *Identifying the poor* (Amsterdam, IOS Press), pp. 35–49.

Kooiman, P., J.G. Bethlehem and J.R. Nobel. Academic and Official Statistics in the Netherlands: Some Perspectives. In: Eurostat (eds.). *Academic and Official Statistics Cooperation* (Luxemburg, European Communities), pp. 59–71.

Kooiman, P., A. van Krimpen and P. Struijs. *Organisation and Management Structures; Report based on the results of a survey among seven national statistical offices: Canada, Australia, New Zealand, Norway, Finland, Sweden, Denmark.*

Maarseveen, J.G.S.J. van. Statistics Netherlands looks back a century. *Sigma, The Bulletin of European Statistics*, 3, pp. 40–41.

Maarseveen, J.G.S.J. van. Statistics Netherlands celebrates its centennial. *Netherlands Official Statistics*, 14, Autumn, pp. 26–29.

Maarseveen, J.G.S.J. van, M.B.G. Gircour and R. Schreijnders (eds.). *A century rounded up. Reflections on the history of the Central Bureau of Statistics in the Netherlands* (Stichting beheer IISG, Amsterdam).

Maarseveen, J.G.S.J. van and M.B.G. Gircour (eds.). *A century of statistics. Counting, accounting and recounting in the Netherlands* (Stichting beheer IISG, Amsterdam).

Vries, W. de. Ranking: right or wrong? Some problems in comparing national statistical offices and systems. *Netherlands Official Statistics*, 14, Autumn, pp. 4–6.

Vries, W.F.M. de and J.R. Nobel. Statistik, Geheimnisse und Empfindingen. In: J. Chlumsky, B. Schimpl-Neimanns and G. Wagner (eds.). *Kooperation zwischen Wissenschaft und amtlicher Statistik – Praxis und Perspektiven* (Wiesbaden, Statistisches Bundesamt), pp. 158–175.

Agriculture, manufacturing and the environment

Abrahamse, A.P.J. and T. Werkhoven. Business cycle surveys for the manufacturing industry. *Netherlands Official Statistics*, Summer, pp. 4–8.

Audretsch, D., L. Klomp and R. Thurik. Do services differ from manufacturing? The post-entry performance of firms in Dutch services. In: *Innovation, Industry Evolution, and Employment*, edited by D.B. Audretsch and A.R. Thurik (Cambridge University Press, Cambridge, United Kingdom), pp. 230–252.

Boer, B. de, R. Hueting and M. de Haan. Green accounting and sustainable national income calculation in the Netherlands. In: *From research to implementation: policy-driven methods for evaluating macro-economic environmental performance* (European Union, DG XII, Brussel).

Carree, M.A, L. Klomp and A.R. Thurik. *Productivity Convergence in OECD Manufacturing Industries.* Working Paper TI 99-065/3 in Tinbergen Institute Discussion Paper Series. Amsterdam/ Rotterdam.

Delany, S., C. Reyes, E. Hubert, S. Phil, E. Rees, L. Haanstra and A.J. van Strien. Results from the International Waterbird Census in the Western Palearctic and Southwest Asia, 1995 and 1996. In: *Wetlands International* 54, Wageningen.

Dietz, E.J. *The Dutch experience with collecting environmental data from local governments.* Paper for the ICLEI's 4th International Expert Seminar on Environmental Management Instruments: Developing an Environmental Accounting Standard for Local Governments, University of Canberra, 21–23 April 1999.

Duuren, L. van, J. Schaminee and E. Weeda. Atlas of plant communities in the Netherlands. *Annali di Botanica*, Vol. LVI-1, pp. 93–100.

Duuren, L. van, J. Schaminée and E. Weeda. Atlas of plant communities in the Netherlands. *Netherlands Official Statistics*, Summer, pp. 34–36.

Eerdt, M.M. van and P.K.N. Fong. The monitoring of nitrogen surplusses from agriculture. *Environmental Pollution* 102, S1, pp. 227–233.

Hueting, R. and B. de Boer. Sustainable National Income According to Hueting. In: H. Verbruggen (eds.). *Interim report on calculations of a sustainable national income according to Hueting's methodology* (Institute for Environmental Studies, Vrije Universiteit, Amsterdam).

Klomp, L. and G. van Leeuwen. The importance of innovation for company performance. *Netherlands Official Statistics*, Winter, pp. 26–35.

Klomp, L. and A.R.Thurik. Job Flows of Firms in Traditional Services. In: *Entrepreneurship, Small and Medium-Sized Enterprises and the Macroeconomy*, edited by Z.J. Acs, B. Carlsson, and C. Karlsson (Cambridge University Press, New York), pp. 310–328.

Lanoy Meijer, N. de. *An additional Innovation Survey by Statistics Netherlands.* Room document: Eight EEA Working Party Meeting on R&D and Innovation Statistics, 22–25 November 1999, Luxembourg, Eurostat.

Lengkeek, W. Land use statistics in the Netherlands. *Netherlands Official Statistics*, 14, Autumn, pp. 21–23.

Lord, E., M.M. van Eerdt and E. Bomans. *Nutrient Balances for grassland systems in the EU.* Paper for the meeting of the working group 'Statistics of the Environment', Joint Eurostat/EFTA Group Meeting, 25–26 February 1999.

Meinen, G.W. *Measuring Capital Stock: it's a nasty job but someone has to do it.* Internet discussion paper for Canberra Group on Capital Stock Statistics.

Niphuis, A.J. *Thresholds and coverage implication of agricultural surveys and censuses.* Paper for the Joint UN/ECE-Eurostat-OECD-FAO Meeting on food and agricultural statistics, June 1999.

Oenema, O., P.C.M. Boers, M.M. van Eerdt, B. Fraters, H.G. van der Meer, C.W.J. Roest, J.J. Schröder and W.J. Willems. Leaching of nitrate form agriculture to groundwater: the effect of policies and measures in the Netherlands. *Environmental Pollution* 102, S1, pp. 471–478.

Strien, A.J. van and J. Pannekoek. A pilot study on Euromonitoring. *Bird Census News* (11) 2, pp. 46–49.

Trade, transport and services

Bakkes, R., B. Kombert-Engelhardt, F. van de Pol and J. Walter. *Reconciliation Exercise of Foreign Trade Statistics in Germany and the Netherlands* (Statistisches Bundesamt, Wiesbaden, and Statistics Netherlands, Heerlen).

Goedegebuure, R. and M. Luppes. *The policy relevance of official statistics: the case of distribution centers in the Netherlands*.

Moritz, G. and W. Brög. Redesign of the Dutch Travel Survey: Response improvement. Paper presented at the Transportation Research Board Conference on Personal Travel: The Long and Short of It (Washington, DC, June–July, 1999). *Netherlands Official Statistics*, 14, Autumn, pp. 7–12.

Pol, F. van de and R. Janssen. *A Dutch View on Intrastat II and III*.

Slootbeek-van Laar, M. *The system of index numbers (1997) of the statistics of International Trade*.

Socio-cultural statistics

Bakker, B.F.M, K.G. Tijdens and J.W. Winkels. Explaining gender wage differences. *Netherlands Official Statistics*, Winter, pp. 36–41.

Cavelaars, A.E.J.M., A.E. Kunst and J.J.M. Geurts. Educational differences in smoking: an international comparison. *British Medical Journal*, 1999, 4.

Faessen, W. Statistics on housing and construction in the Netherlands. Housing demand among the elderly. *Netherlands Journal of Housing and the Built Environment*, 1999, 4, p. 323.

Gringhuis, G.H. and A.Z. Israëls. Analysing well-being in relation to characteristics of the population. *Netherlands Official Statistics*, Summer, pp. 28–33.

Harmsen, E. and J.W. Winkels. Co-ordination of population and household characteristics. In: J.G.S.J. van Maarseveen and M.B.G. Gircour (eds.), *A century of statistics. Counting, accounting and recounting in the Netherlands* (Stichting beheer IISG, Amsterdam).

Kempkens, C.M.O.A. *Integrated system of social surveys on living conditions*. Paper presented at the GSS Methodological Conference, 21 June 1999, London.

Leiden, H.A. van, A.A.M.W. van Gessel-Dabekaussen, P.J. van der Maas and H.J. de Koning. Trends in mammography 1991–96 and the impact of nationwide screening in the Netherlands. *Journal of Medical Screening*. 1999, 6, pp. 94–98.

Mikulic, B. and J. Oudhof. Housing situation and low income in the Netherlands. *Netherlands Journal of Housing and the Built Environment*, 1.

Mosseveld, C. *International Comparison of Health Care data in six countries*. Paper presented at the International Symposium on National Health Accounts in the framework of the 2nd international Health Economics Association Congres in Rotterdam, June 1999.

Mosseveld, C.J.P.M. and P. van Son. *International Comparison of Health Care data: methodology development and application* (Kluwer Academic Publishers, Dordrecht).

Swinkels, H., B. Jonas and J. v.d. Berg. *The correlation between physical health and mental health in the Netherlands and the US*.

Socio-economic statistics

Alders, M. and D. Manting. *Household scenarios for the European Union, 1995–2025*. Paper presented at the European Population Conference, 30 August – 3 September, The Hague.

Alders, M. and D. Manting. Household scenarios for the European Union: methodology and main results. *Netherlands Official Statistics*, Summer, pp. 17–27.

Beer, J. de. Demographic trends in the 21st century. In: J. Garssen et al. (eds.), *Vital events. Past, present and future of the Dutch population*, pp. 123–139.

Beer, J. de and M. Alders. *Probabilistic population and household forecasts for the Netherlands*. Paper presented at the Joint ECE/Eurostat Work Session on Demographic Projections, 3–7 May, Perugia and at the European Population Conference 1999, 30 August–3 September, The Hague.

Beer, J. de and M. Alders. Uncertainty of population forecasts: a stochastic approach. *Netherlands Official Statistics*, Winter, pp. 19–25.

Beer, J. de, R. Broekman, N. van der Gaag, W. van Hoorn, C. Huisman, A. de Jong, M. Mellens and L. van Wissen. *New population scenarios for Europe: uniformity or diversity?* Paper presented at the Joint ECE/Eurostat Work Session on Demographic Projections, 3–7 May, Perugia

Beer, J. de, and L. van Wissen. One Europe: how many different worlds in the 21st century? In: J. de Beer and L. van Wissen (eds.), *Europe: one continent, different worlds. Population scenarios for the 21st century* (Kluwer Academic Publishers, Dordrecht), pp. 149–161.

Berkel, K. van, and J. van der Valk. Restructuring the Dutch Labour Force Survey. *Netherlands Official Statistics*, 14, Autumn, pp. 18–20.

Bernelot Moens, M. The Dutch standard classification of education. *Netherlands Official Statistics*, 14, Autumn, pp. 24–25.

Buhmann, B., W.P. Leunis, A.Vuille and K. Wismer. *Labour accounts principles and practice, experiences in Denmark, the Netherlands and Switzerland*. Paper presented at a seminar on labour accounts in Copenhagen, 14 December 1999.

Driel, J. van. *Estimating annual numbers of courses and students from the labour force surveys in the EEC*. Research paper.

Graaf, A. de and A.H. Sprangers. Older mothers, fewer children. In: J. Garssen et al. (eds.*), Vital events. Past, present and future of the Dutch population*, pp. 51– 65.

Haan, J. de. *Demographic Change and Upper Level Bias in a 'Democratic' CPI*. Paper presented at the International Conference on 'The Measurement of Inflation', Cardiff, 31 August–1 September.

Haan, J. de. *Dealing with New and Disappearing Goods in the Fisher Price Index*. Research paper.

Haan, J. de. Linking Price Index Weights to National Accounts Expenditure Data. *Research in Official Statistics*, 2, pp. 75–89.

Haan, J. de, E. Opperdoes and C. M. Schut. Item Selection in the Consumer Price Index: Cut-off versus Probability Sampling. *Survey Methodology*, 25, pp. 31–41.

Harmsen, C.N. and K. Prins. A rapid increase in numbers. In: J. Garssen et al. (eds.), *Vital events. Past, present and future of the Dutch population*, pp. 101–120.

Harmsen, C.N. *Cross-cultural marriages*. Paper presented at the European Population Conference 1999, 30 August – 3 September 1999, The Hague.

Hoorn, W. van, and J. Garssen. The cautious retreat of death. In: J. Garssen et al. (eds.), *Vital events. Past, present and future of the Dutch population*, pp. 85–99.

Hoorn, W. van, and R. Broekman. Uniformity and diversity scenarios for mortality. In: J. de Beer and L. van Wissen (eds.), *Europe: one continent, different worlds. Population scenarios for the 21st century* (Kluwer Academic Publishers, Dordrecht), pp. 71–90.

Hoorn, W. van, N. van der Gaag and C. Huisman. Population development in Europe in the 21st century. In: J. de Beer and L. van Wissen (eds.), *Europe: one continent, different worlds. Population scenarios for the 21st century* (Kluwer Academic Publishers, Dordrecht), pp. 109–147.

Hoorn, W. van. *Glad to live alone or happier together? Diversity among young and middle-aged single people*. Paper presented at the European Population Conference 1999, 30 August–3 September 1999, The Hague.

Huis, L.T., H. Nicolaas and M. Croes. *Migration of the four largest cities in the Netherlands*. Paper presented at the European Population Conference 1999, 30 August–3 September 1999, The Hague.

Jong, A. de. *Population and labour force scenarios for the European Union: acceleration, continutity or reversal.* Paper presented at the European Population Conference 1999, 30 August–3 September 1999, The Hague.

Jong, A. de. Labour force scenarios for the European Union. *Netherlands Official Statistics*, Winter, pp. 10–18.

Jong, A. de, and R. Broekman. Uniformity and diversity scenarios for fertility. In: J. de Beer and L. van Wissen (eds.), *Europe: one continent, different worlds. Population scenarios for the 21st century* (Kluwer Academic Publishers, Dordrecht), pp. 45–69.

Jong, A. de, and A. de Graaf. Marriage: from cornerstone to outdated institution? In: J. Garssen et al. (eds.), *Vital events. Past, present and future of the Dutch population*, pp. 37–49.

Jong, A. de, and W. van Hoorn. Leaving home, and then? In: J. Garssen et al. (eds.), *Vital events. Past, present and future of the Dutch population*, pp. 21–34.

Laan, P. van der. *Inventory of the Main Registers and Administrative Sources being used for Social and Demographic Statistics in the Netherlands*. Supporting Paper prepared for the Joint ECE-Eurostat Work Session on Registers and Administrative Records in Social and Demographic Statistics. Working Paper No. 22. United Nations, Economic and Social Council, Statistical Commission and Economic Commission for Europe, Conference of European Statisticians, and Statistical Office of the European Communities. Geneva, Switzerland, 1–3 March.

Laan, P. van der. *Major Issues in Producing Income Statistics*. Paper prepared for the Seventh Seminar of the European Advisory Committee on Statistical Information in the Economic and Social Spheres, 'Income Distribution and Different Sources of Income'. Cologne, Germany, 10–11 May.

Laan, P. van der. *The Problematic Measurement of Income from Self-Employment*. Paper prepared for the Eurostat Seminar on Income Methodology for Statistics on Households. Statistical Office of the European Communities. Luxembourg, 13–14 December.

Leunis, W.P. *Linking social and economic statistics through the 1995 revision of National accounts and labour accounts*. Paper presented at the OECD national accounts working party meeting, 21–24 September 1999, Paris.

Leunis, W.P. and S. Keuning. *Statistical integration, the accounting approach; Challenges of combining economic and social statistics in an accounting system*.

Mellens, M. Determinants of demographic behaviour. In: J. de Beer and L. van Wissen (eds.*), Europe: one continent, different worlds. Population scenarios for the 21st century* (Kluwer Academic Publishers, Dordrecht), pp. 5–32.

Mellens, M. Uniformity and diversity defined. In: J. de Beer and L. van Wissen (eds.*), Europe: one continent, different worlds. Population scenarios for the 21st century* (Kluwer Academic Publishers, Dordrecht), pp. 33–44.

Nicolaas, H. *Family-network migration after asylum migration in the Netherlands*. Paper presented at the European Population Conference 1999, 30 August – 3 September 1999, The Hague.

Nicolaas, H., A.H. Sprangers and J. Garssen. Immigrants outnumber emigrants. In: J. Garssen et al. (eds.), *Vital events. Past, present and future of the Dutch population*, pp. 67–83.

Schulte Nordholt, E. Statistical disclosure control of Statistics Netherlands employment and earnings data. *Netherlands Official Statistics*, Spring, special issue on statistical disclosure control, pp. 34–38.

Schulte Nordholt, E. The role of metadata in the statistical production process. In: *Bulletin of the International Statistical Institute*, 52nd Session, Helsinki, August 10–18, Proceedings, Tome LVIII, Book 3, Helsinki, Finland, 1999, pp. 237–238.

Schulte Nordholt, E. Imputation, the alternative for surveying earning patterns. In: *Official Statistics in a Changing World, Proceedings of the 3rd International Conference on Methodological Issues in Official Statistics*, October 12–13, 1998, Statistics Sweden, Stockholm, Sweden, pp. 159–162.

Schulte Nordholt, E. *Statistical Disclosure Control of the Statistics Netherlands employment and earnings data*. Presentation at the Joint CE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, March.

Schulte Nordholt, E. *Statistical Disclosure Control: from rools to tools*. Presentation at the Seminar om konfidensialitet, Oslo, June.

Schulte Nordholt, E. *Hot deck imputation in the Dutch Structure of Earnings Survey*. Presentation at the International Seminar on Exchange of Technology and Know-how ETK, Prague, October

Schulte Nordholt, E. *Automatic editing in the Dutch Labour Cost Survey using Cherrypi*. Research paper prepared for UN/ECE, Work Session on Statistical Data Editing, Rome

Schulte Nordholt, E. *The role of metadata in the statistical production process*. Research paper prepared for the 52nd Session of the International Statistical Institute in Helsinki.

Sprangers, A.H. *Fertility of foreign-born women in the Netherlands*. Paper presented at the European Population Conference 1999, 30 August–3 September 1999, The Hague.

Sprangers, A.H. and J.B. Sanders. *Now-casts on international migration – Part 1: creation of an information database*. Eurostat Working Paper.

Statistics Netherlands. *Labour accounts in theory and practice; The Dutch experience*.

Statistics Netherlands. *The development of quarterly estimations of labour cost for the Netherlands, Methodology*. Final report. Working paper prepared on request of Eurostat.

Verhoef, R. and J. Garssen. Why demography matters. In: J. Garssen et al. (eds.), *Vital events. Past, present and future of the Dutch population*, pp. 7–18.

Vliegen, J.M. and P. van der Laan. Der 'Zensus' in den Niederlanden: Eine Integration von Register- und Stichproben-daten'. In: Volkszählung 2001: Von der traditionellen Volkszählung zum Registerzensus, hrsg. H. Grohmann, H. Sahner und R. Wiegert, S. 15–23. *Sonderhefte zum Allgemeinen Statistischen Archiv*, Heft 33. Göttingen: Vandenhoeck & Ruprecht.

Vliegen, J.M. and P. van der Laan. Methodische und zeitliche Aspekte der Umstellung der amtlichen Statistik auf Register am Beispiel der Niederlande'. *Allgemeines Statistisches Archiv*, 83. Band, Heft 4 (Oktober–Dezember 1999), pp. 434–446.

Presentation and integration

Algera, S. Structural Business Statistics: a key role in the system of economic statistics, the Dutch case. *Netherlands Official Statistics*, Winter, pp. 4–9.

Vliegen, J.M. and N. van Leeuwen. Digital topographical maps as a source for the compilation of area-based statistics. *Netherlands Official Statistics*, Summer, pp. 13–16.

Data collection

Akkerboom, J.C. and Håkan L. Lindstrøm. *The effects of data collection characteristics on employment and unemployment estimates*. Final report of Tender No. 97/S 112-69506/EN

Heer, W. de. International response Trends: Results of an International Survey. *Journal of Official Statistics*, 1999, vol. 15, no. 2.

Heer, W. de and E. de Leeuw. Surveying individuals and households, a passionate history. In: J.G.S.J. van Maarseveen and M.B.G. Gircour (eds.), *A century of statistics. Counting, accounting and recounting in the Netherlands* (Stichting beheer IISG, Amsterdam).

Heer, W. de, E. de Leeuw and J. van der Zouwen: Methodological Issues in Survey Research: A historical Overview. *Bulletin de Methodologie Sociologique*, Oct. 1999, n. 64.

Kloek, W., and J. Ritzen. *Sources for co-ordinated population*. Paper for the 13th International Roundtable on Business Survey Frames.

Kloek, W. *The improvement of the business register for statistical purposes.* Report.

Kloek, W. The development of data collection from businesses. In: J.G.S.J. van Maarseveen and M.B.G. Gircour (eds.), *A century of statistics. Counting, accounting and recounting in the Netherlands* (Stichting beheer IISG, Amsterdam).

Lammers, J. *'Maintaining business register, links with administrative register' on behalf of Caucasian countries.* Program for TACIS-action AGA 99001. CBS, Heerlen, 1–5 February 1999.

Snijkers, G., J. Hox and E.D. de Leeuw. Interviewers' Tactics for Fighting Survey Non-response. *Journal of Official Statistics*, vol. 15, nr. 2, pp. 185–198.

Snijkers, G., E. de Leeuw, D. Hoezen and I. Kuijpers. Computer-Assisted Qualitative Interviewing: Testing and Quality Assessment of CAPI and CATI Questionnaires in the field. In: Banks, R., et al. (eds.). *Leading Survey and Statistical Computing into the New Millennium. Proceedings of the Third ASC International Conference*, pp. 231–258 (The University of Edinburgh, Scotland, UK).

Struijs, P. *Business Registers and Structural Business Statistics in the Netherlands.* Paper presented at the Seminar on Structural Business Statistics in Warsaw, Poland, 7–9 June 1999.

Struijs, P. *Metadata at Statistics Netherlands.* Paper presented at the Research Conference van de Federal Committee on Statistical Methodology in Arlington, Virginia, 15–17 November 1999.

Struijs, P. *Tracking Real Changes in Business Structures: a Conceptual Framework.* Paper presented at the INSEE seminar Changing for the Better: Approaches to Restructuring Corporate Groups.

Research and development

Waal, T. de, R. Renssen and F. van de Pol, 1999, Graphical Macro Editing: Possibilities and Pitfalls.

Brakel, J.A. van den and R.H. Renssen. Testing effects of incentives and a condensed questionnaire on response rates. *Netherlands Official Statistics*, Summer, pp. 9–12.

Huisman, M. Relationships between registered unemployed and response rates in the Labour Force Survey. *Netherlands Official Statistics*, 14, Autumn, pp. 13–17.

Kooiman, P., J. Nobel and L. Willenborg. Statistical data protection at Statistics Netherlands. *Netherlands Official Statistics*, Spring, special issue on statistical disclosure control, pp. 21–25.

Vucsan, M.H.J. The application of data warehousing techniques in a statistical environment.

Waal, T. de and L. Willenborg. Exact disclosure in a super-table. *Netherlands Official Statistics*, Spring, special issue on statistical disclosure control, pp. 11–16.

Waal, T. de and L. Willenborg. Information loss through global recoding and local suppression. *Netherlands Official Statistics*, Spring, special issue on statistical disclosure control, pp. 17–20.