Research on the quality of registers to make data decisions in the Dutch Virtual Census

Eric Schulte Nordholt[1]
Saskia J.L. Ossen[2]
Piet J.H. Daas[3]

Abstract

Since the last Census based on a complete enumeration was held in 1971, the willingness of the population in the Netherlands to participate has decreased tremendously. Statistics Netherlands found an alternative in a Virtual Census, by using available registers and surveys as alternative data sources. Advantages of a Virtual Census are that it is cheaper and more socially acceptable. The combined use of registers and surveys for composing the Census however also leads to several methodological challenges. One of them is determining the effect of the quality of the sources. For registers, for instance, the collection and maintenance is beyond the control of the Statistical Agency. It is therefore important that the Statistical Agency is able to determine the quality of the sources used. Insight into the quality of the sources used enables a well thought-out comparison between comparable information in various sources.

Keywords: Quality of registers, Checklist, Census.

Investigación sobre la calidad de los registros para elaborar decisiones de datos en el Censo Virtual Holandés

Resumen

Desde que el último Censo basado en la enumeración universal se realizó en los Países Bajos en 1971, la predisposición de los holandeses a participar en el Censo

[1] Statistics Netherlands, Division of Socio-economic and spatial statistics. (e.schultenordholt@cbs.nl).

[2] Statistics Netherlands, Division of Process development, IT and methodology (sjl.ossen@cbs.nl).

[3] Statistics Netherlands, Division of Process development, IT and methodology (pjh.daas@cbs.nl).

Fecha de recepción: diciembre 2011. Fecha de aceptación: febrero 2013

ha disminuido notablemente. El Instituto de Estadística Holandés diseñó como fórmula alternativa al Censo tradicional el Censo Virtual, basado en la combinación de registros y encuestas. Las ventajas del Censo Virtual son su bajo coste y su mayor aceptación social. El uso combinado de registros y encuestas para generar un Censo Virtual plantea algunos retos metodológicos importantes. Uno de ellos es examinar la calidad de las fuentes. En el caso de los registros, por ejemplo, la recogida y mantenimiento de los mismos está fuera de control del Instituto de Estadística. En consecuencia, es importante que el Instituto de Estadística pueda examinar la calidad de las fuentes utilizadas para examinar la coherencia de la información entre varias fuentes.

Palabras claves: Calidad de los registros, censos.

Recherche sur la qualité des registres pour élaborer des décisions de données dans le Recensement Virtuel Hollandais

Résumé

Depuis le dernier Recensement basé sur l'énumération universelle, réalisé en 1971 aux Pays-Bas, la volonté des Néerlandais à y participer a diminué nettement. L'Institut de Statistiques Néerlandais a dessiné, comme formule alternative au Recensement traditionnel, un Recensement Virtuel qui s'appuie sur la combinaison de registres et d'enquêtes. Les avantages du Recensement Virtuel sont son faible coût et une meilleure tolérance sociale. L'usage combiné de registres et d'enquêtes afin de générer un Recensement Virtuel pose néanmoins quelques défis méthodologiques importants. L'un d'eux est examiner la qualité des sources. Dans le cas des registres, par exemple, la collecte et l'entretien de ces registres échappe au contrôle de l'Institut de Statistiques. Par conséquent, il est important que l'Institut de Statistiques puisse examiner la qualité des sources utilisées de manière à garantir la cohérence de l'information issue de diverses sources.

Mots clés: Qualité des registres, recensements.

INTRODUCTION[4]

All European Union (EU) countries will conduct a Census in 2011. The way this Census will be conducted is up to the countries. In the Netherlands virtual censuses are held ever since the last traditional

[4] The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. This paper has been evaluated anonymously.

Census in 1971. This means that census forms no longer exist and that the relevant information is provided by data in already existing registers and surveys (Schulte Nordholt, 2004). In this way the Virtual Censuses of 1981, 1991, and 2001 were conducted. The Censuses of 1981 and 1991 were of a limited character. The data compiled on 1981 and 1991 were much less detailed than the set of tables of the 2001 Census. In 2001 Statistics Netherlands published census information on the municipal level. For the 2011 Census even more registers and surveys will be combined. The Population Register forms the backbone for the integration activities that will eventually result in coherent and detailed demographic and socio-economic statistical information on persons and households.

A generic problem in using administrative registers for statistical purposes is that the data in these sources are collected and maintained by other organizations for non-statistical purposes. The process is beyond the control of Statistics Netherlands. This not only makes Statistics Netherlands highly dependent, it may also affect the quality of the output of Statistics Netherlands. As Statistics Netherlands is expected to use more and more registers in the future in order to lower the administrative burden, a quality framework has been developed that enables the determination of the quality of externally collected data sources, such as registers, prior to use (Daas et al., 2009). This framework was used to study the input quality of the most important registers used in the Virtual Census 2011. The results of these studies are the topic of this paper. The Nordic countries are in a similar situation with their censuses in the sense that more and more registers are being used (United Nations, 2007). In the following section the data sources and variables of the 2011 Census in the Netherlands considered in this paper are introduced. In section 3 the quality framework is described in more detail. Next the results of applying the framework are discussed. Finally, some conclusions are drawn in section 5.

1. DATA SOURCES AND VARIABLES

The Population Register (PR) is the backbone of the Census. Information from other registers and surveys is added to eventually derive all 2011 Census variables. It is important to realize that registers change over time and so does their quality.

For example, the new Housing Register (HR) was not yet available for the 2001 Census but is going to be used in the 2011 Census. It is to be expected that part of the information in the new HR is able to replace information that -in the 2001 Census project- was provided by two other data sources; viz. the old Housing Register and the Survey on Housing Conditions (SHC).

In addition, the fiscal and social security registers in the Netherlands have also changed since the 2001 Census. These data sources have merged and will be used instead of the formerly used Survey on Employment and Earnings (SEE). It is our hope that this new combined register, together with the Unemployment Benefit Register (UR) and the Social Security Register (SR), can be used to derive most categories of the variable current activity status. All persons who are not employed and appear in the UR or SR are then considerd to be unemployed. In addition to register information, some information provided by the Labour Force Survey (LFS) remains essential for the 2011 Census. All information on persons can be easily combined based on the unique Citizen Service Number.

The decisions about which data sources are used to produce the different variables in the 2011 Census are predominantly based on the quality of the sources containing information about the variables. In this paper a number of registers will be compared for a limited set of three variables. These are: highest level of educational attainment, current activity status, and housing information.

The highest level of educational attainment is an important variable. Information regarding this variable can be found in the LFS. Nevertheless, the Dutch LFS contains only a small fraction (approximately 1 %) of the population aged 15-64 per calendar year. Information about many more people can be found in the Education Register (ER). However, the information in the Dutch ER is less recent than in the LFS. Ideally, information from both sources is combined. For the Census, information from one of these sources might be enough to produce reliable consistent tables.

Current activity status is in fact a variable that includes many different categories as e.g. employed, unemployed and homemakers. Information about employed people comes from register information. Information about unemployment according to the International Labour Organization (ILO) definition can be obtained on the basis of LFS survey data. Another option is to derive unemployment from register information containing benefits: viz. the UR and the SR. The information in these registers is integral but does not have the exact definition of unemployment needed for the Census. The research question here is what information is best for the 2011 Census: sample information from the LFS with the correct definition or integral information from registers with an approximation of the official definition?

Housing information can be obtained from the new HR. As stated before, this register has not been used for earlier censuses. A disadvantage of this register is that it lacks some information. Since some of the variables in the HR are also available in other sources (e.g. in the land register), the question is which of the sources should be used to derive specific Census variables.

The brief overview given above clearly reveals that the sources ER, UR, SR, HR, and PR all provide useful information for deriving one of the variables under concern. In this paper the current state and quality of the information about level of education, current activity status, and housing available in the registers (and in the LFS) will be studied using the quality framework for registers.

2. QUALITY FRAMEWORK

The quality framework for registers was developed to standardize the determination of the various quality components of administrative registers (Daas et al., 2009). The quality framework consists of three high level views on quality. These three high level views give a complete overview of the quality components (Daas et al., 2010). These views are referred to as hyper dimensions (Karr et al., 2006) and are called: Source, Metadata, and Data. Each hyper dimension is composed of several dimensions of quality and each dimension contains a number of quality indicators. A quality indicator is measured or estimated by one or more methods which can be qualitative or quantitative (Daas et al., 2009). Subsection 3.1 starts with an overview of the quality aspects in the Source and Metadata hyper dimension and the methods developed to determine them. Next, recent insights on the study of the quality aspects in the Data hyper dimension are described (Daas et al., 2011).

2.1. Source and Metadata hyper dimensions

A statistical office that plans to use an administrative register should start by exploring the quality of the information that enables the use of the data source on a regular basis. These components of quality are located in the Source hyper dimension of the quality framework. In table 1 the dimensions, quality indicators, and method descriptions for this hyper dimension are shown. The second hyper dimension in the framework, the Metadata hyper dimension, focuses on the conceptual and process related quality components of the metadata of the source. Prior to use, it is essential that a statistical office fully understands the metadata related quality components because any misunderstanding highly affects the quality of the output based on the data in the source. In table 2 the dimensions, quality indicators, and method descriptions are shown for the Metadata hyper dimension.

TABLE 1
Quality framework for secondary data sources. Source hyper dimension

Dimensions	Quality indicators	Methods
1. Supplier	1.1 Contact	-Name of the data source -DSH[1] contact information -NSI[2] contact person
	1.2 Purpose	-Reason for use of the data source by DSH
2. Relevance	2.1 Usefulness	-Importance of data source for NSI
	2.2 Envisaged use	-Potential statistical use of data source
	2.3 Information demand	-Does the data source satisfy information demand?
	2.4 Response burden	-Effect of data source on response burden
3. Privacy & security	3.1 Legal provision	-Basis for existence of data source
	3.2 Confidentiality	-Does the Personal Data Protection Act apply? -Has use of data source been reported by NSI?
	3.3 Security	-Manner in which the data source is send to NSI -Are security measures required? (hard-/software)
4. Delivery	4.1 Costs	-Costs of using the data source
	4.2 Arrangements	-Are the terms of delivery documented? -Frequency of deliveries
	4.3 Punctuality	-How punctual can the data source be delivered? -Rate at which exceptions are reported -Rate at which data is stored by DSH
	4.4 Format 4.5 Selection	-Formats in which the data can be delivered -What data can be delivered? -Does this comply with the requirements of NSI?
5. Procedures	5.1 Data collection	-Familiarity with the way the data is collected
	5.2 Planned changes	-Familiarity with planned changes of data source -Ways to communicate changes to NSI
	5.3 Feedback	-Contact DSH in case of trouble? -In which cases and why?
	5.4 Fall-back scenario	-Dependency risk of NSI -Emergency measures when data source is not delivered according to arrangements made

[1] DSH: Data Source Holder; [2] NSI: National Statistical Institute.

For the evaluation of the quality indicators in the Source and Metadata hyper dimension a checklist has been developed. It is included in the paper of Daas et al. (2009). The checklist guides the user through the measurement methods for each of the quality indicators in both hyper dimensions. By answering the questions in the checklist, the 'value' of every method for each indicator in tables 1 and 2 is determined, ranging from good to poor. Evaluation of the Metadata-part requires that the user has a particular use in mind, which is the 2011 Census in our case. The next step is the determination of the quality of the data.

TABLE 2
Quality framework for secondary data sources, Metadata hyper dimension

Dimensions	Quality indicators	Methods
1. Clarity	1.1 Population unit definition	-Clarity score of the definition
	1.2 Classification variable definition	-Clarity score of the definition
	1.3 Count variable definition	-Clarity score of the definition
	1.4 Time dimensions	-Clarity score of the definition
	1.5 Definition changes	-Familiarity with occurred changes
2. Comparability	2.1 Population unit definition comp.	-Comparability with NSI definition
	2.2 Classification variable def. comp.	-Comparability with NSI definition
	2.3 Count variable definition comp.	-Comparability with NSI definition
	2.4 Time differences	-Comparability with NSI reporting periods
3. Unique keys	3.1 Identification keys	-Presence of unique keys -Comparability with unique keys used by NSI
	3.2 Unique combinations	-Presence of useful combinations of variables
4. Data treatment (by DSH)	4.1 Checks	-Population unit checks performed -Variable checks performed -Combinations of variables checked -Extreme value checks
	4.2 Modifications	-Familiarity with data modifications -Are modified values marked and how? -Familiarity with default values used

2.2. Data hyper dimension

Indicators for the evaluation of the quality of the data in a register are part of the Data hyper dimension. The focus of the indicators in this dimension is the quality of the data in the registers used as input in the statistical process (Daas et al., 2011). The indicators and dimensions identified are listed in table 3.

TABLE 3
Quality framework for secondary data sources, Data hyperdimension

Dimensions	Quality indicators	Methods
1. Technical checks	1.1 Readability	-Accessability of the file and data in the file
	1.2 File declaration	-Compliance of the data to the metadata agreements
	1.3 Convertibility	-Conversion of the file to the NSI-standard format
2. Accuracy		Objects
	2.1 Authenticity	-Legitimacy of objects
	2.2 Inconsistent objects	-Extent of erroneous objects in source
	2.3 Dubious objects	-Presence of untrustworthy objects
		Variables
	2.4 Measurement error	-Deviation of actual data value from ideal error-free measurements
	2.5 Inconsistent values	-Extent of inconsistent combinations of variable values
	2.6 Dubious values	-Presence of implausible values or combinations of values for variables
3. Completeness		Objects
	3.1 Undercoverage	-Absence of target object in the source
	3.2 Overcoverage	-Presence of non-target objects in the source
	3.3 Selectivity	-Statistical coverage and representativiness of objects
	3.4 Redundancy	-Presence of multiple registrations of objects
		Variables
	3.5 Missing values	-Absence of values for (key) variables
	3.6 Imputed values	-Presence of values resulting from imputation
4. Time-related dimension	4.1 Timeliness	-Time between end of reference period and receipt of source
	4.2 Punctuality	-Time lag between the actual and agreed delivery date
	4.3 Overall time lag	-Overall time difference between end of reference period and moment NSI concluded source can be used
	4.4 Delay	-Extent of delays in registration
		Objects
	4.5 Dynamics of objects	-Changes in the population of objects over time
		Variables
	4.6 Stability of variables	-Changes of variables or values over time
5. Integrability		Objects
	5.1 Comparability of objects	-Similarity of objects in source with the NSI-objects
	5.2 Alignment of objects	-Linking-ability of objects in source with NSI-objects
		Variables
	5.3 Linking variable	-Usefulness of linking variables (keys) in source
	5.4 Comparability of variables	-Proximity (closeness) of variables

3. QUALITY EVALUATION RESULTS

The checklist referring to the Source and Metadata hyper dimension has been applied to the aforementioned registers. Next to that a first step has been made in applying the indicators corresponding to the Data hyper dimension. In this section first the evaluation results of applying the checklist to the various registers are discussed. Next some preliminary findings of the quality evaluation regarding the Data hyper dimension are presented. The focus of this study was the level of education, the current activity status, and housing information available in the registers.

3.1. Source and Metadata: application of checklist

The checklist was applied to the ER, UR, SR, HR, and PR registers. The evaluation results obtained for the Source and Metadata hyper dimensions are shown in tables 4 and 5, respectively.

TABLE 4
Evaluation results for the Source hyper dimension

Dimensions	Data sources				
	ER	UR	SR	HR	PR
1. Supplier	+	o	o	+	+
2. Relevance	o	+	+	o	+
3. Privacy and security	+	+	+	+	+
4. Delivery	-	+	+	+	+
5. Procedures	o	o	o	+	+

TABLE 5
Evaluation results for the Metadata hyper dimension

Dimensions	Data sources				
	ER	UR	SR	HR	PR
1. Clarity	+	+	+	+	+
2. Comparability	-	o	o	+	+
3. Unique keys	+	+	+	o	+
4. Data treatment	+	+	+	+	+

In both tables evaluation scores are indicated at the dimension level. The dimensional scores were obtained by selecting the most commonly observed score for every measurement method in each dimension. The symbols for the scores used are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combining symbols with a slash (/) as a separator.

The results in table 4 reveal that on a dimensional level, the overall scores for the majority of the data sources are quite good in Source. The ER is an exception, here a poor score is observed for delivery. This is the result of the low frequency of delivery (not more often than once a year). The ER also has only a reasonable (o) score for relevance because this source does not satisfy all information demands for the Census. This register suffers severely from selective undercoverage (see next subsection). The UR and SR score only reasonable for supplier and procedures because of the sometimes problematic unclear purpose for the data provider and the high dependency risk of Statistics Netherlands. The HR has a reasonable score for relevance because this source does not satisfy all information demands; it is missing some variables (e.g. whether the dwelling is owned or rented). The PR only has good scores.

The results in table 5 reveal that on a dimensional level, the overall scores for the data sources are also quite good for most dimensions in the Metadata hyper dimension. The clarity and data treatment dimensions show only good results. Again the ER is the only data source with a poor score. This data source scores poor on comparability because the time period variables cannot be transformed easily to the time points used by Statistics Netherlands. The HR only has a reasonable score for unique keys because of the difficult comparability of the unique keys used in this source. This considerably hinders combining this data source with the other sources of information. The UR and SR have reasonable scores for comparability because of time differences in the reporting periods. Positive exception to all of this is again the PR which only has good scores.

Overall the evaluation results for the five data sources reveal that attention should be paid to the supplier, relevance, procedures, and comparability related quality aspects. The results for the PR demonstrate that it is possible to have every quality aspect in the Source and Metadata hyper dimension under control. For the other data sources it can be argued that the results suggest that one or more of the quality aspects in both hyper dimensions require attention.

It was concluded that not many problems were found for using the registers in the Census 2011.

3.2. Data: first evaluation results

In this section preliminary results of applying the indicators referring to the data hyper dimension are discussed. In the available dataset raw data were already pre-processed to a limited extent and linked to the PR. All data furthermore referred to the same date: January 1, 2008. This implies that the indicators referring to the dimensions: Technical checks, Time-related, and Integratibility are not considered in this paper. The analysis therefore rather focuses on the Completeness and the Accuracy dimension.

3.2.1. Completeness dimension

The analysis on completeness will concentrate on the variables: educational attainment (derived from the ER), and current economic activity status (derived from the UR, SR, and LFS). The housing variables for the 2011 Census all come from registers that are complete and are therefore not discussed further in this paper.

To get a first impression of the level of undercoverage of the information available regarding these variables we assume that the population consists of all persons in the PR. As the data used contains a row for every person in the PR, we consider for how many rows a value is missing. This can be misleading of course when variables are studied that are only applicable to a part of the population. This does however not apply to the variables considered here as the categories of these variables are such that every person belongs to a category. Table 6 provides an overview of the number of missing values for both variables studied.

TABLE 6
Overview of missing values

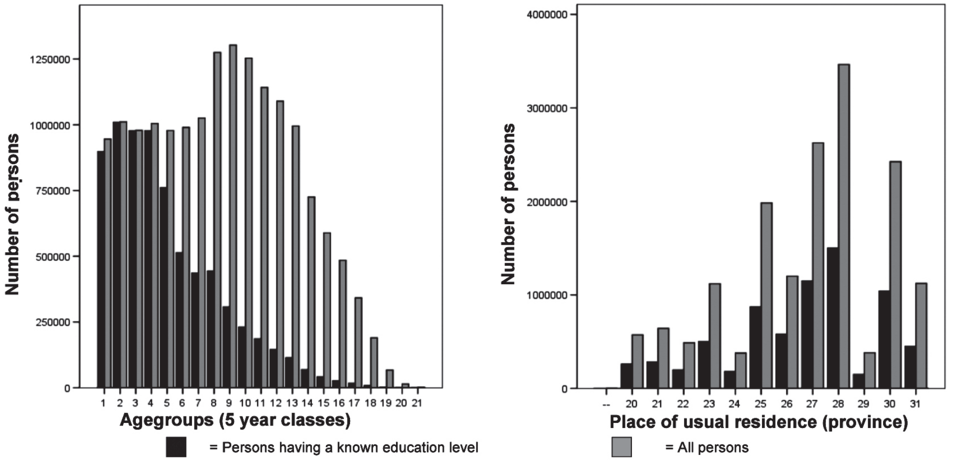
Variable	Number of missings	Percentage missing (%)
Educational attainment	9.238.212	56,3
Current activity status	2.140.266	13,0

Source: own calculations

The undercoverage regarding the variable current activity status is not surprising as for three categories of this variable (i.e. unemployed, homemakers and others) only information from the LFS (based on samples) is present in our data. More serious undercoverage exists for the variable educational attainment. This is also not unexpected as the registers used for deriving this variable contain mainly information about people under the age of 45. This knowledge suggests that the undercoverage is seriously selective regarding age. The selectivity of the information available regarding educational attainment is further examined in figure 1. The blue bars in the histograms refer to the part of the population having a known value for the variable educational attainment. The green bars refer to the population as a whole. In the histogram on the left, people are grouped into age groups (5 year classes), while the histogram on the right groups people based on the variable “Place of usual residence” (provinces).

The figure clearly shows that the available information regarding educational attainment is seriously selective regarding age, i.e. much more information is available about the lower age groups. On the other hand, the shares of people living in the 12 provinces show a strong resemblance between the whole population and the part of the population for which information about educational attainment is available. This suggests that the information is not selective regarding the place where people live.

FIGURE 1
Educational attainment per age group reveals undercoverage for some age groups



Another indicator regarding completeness is redundancy, i.e. the presence of multiple registrations of objects. To investigate whether or not data suffered from redundancy, we searched for rows in our dataset which showed equal values for all variables (including educational attainment and current activity status) except for person_id. There turned out to be 67.644 "duplicates" in our test data, corresponding to 0,4% of the data. A further analysis of the duplicates revealed that most duplicated records corresponded to people living in institutions. People living in homes for the elderly, for example, do all have the same address, are all in the same age category and so on. Given that it is possible that people in institutions do have the same values for the limited set of variables available in our test database, we concluded to focus in future research at the selective part of duplicates not corresponding to people living in institutions.

3.2.2. Accuracy dimension

Regarding the accuracy dimension we consider in this subsection whether there are any dubious values in the data. We concentrate again on the variables educational attainment and current activity status. For these variables especially the relation with age is interesting. For illustration, reaching the highest levels of education takes time. Therefore, for example, a person of 18 years old can (normally) not yet have a PhD degree. Furthermore, it is expected that (almost) only elderly people will have a value for the variable current activity status equal to 3 (pension or capital income recipients).

To be able to analyse whether these expected relations between the variables of interest and age hold, cross tabulations were created. The results are shown in table 7 and table 8. In interpreting these results, care has to be taken of the fact that especially for the variable educational attainment a lot of values are missing and that the number of missings depends on age (see the first column of table 7). Because of this, not much can be concluded from, for example, the counts per cell of the table. Despite of this, for the variable educational attainment it is valid to conclude that for the youngest part of the Dutch population either no value is present or it is equal to "not applicable". The youngest people that have reached education level 6 (Second stage of tertiary education) are in the age group 20-24. Furthermore, there turn out to be some people who reached educational level 5 (First stage of tertiary education) already within the age group 15-19. Most people reach this level however at a higher age. It can also be cautiously concluded that most young people continue studying after they have reached level 1. This is in line with the expectations as youngsters are obliged to go to school till the age of 16 in the Netherlands.

TABLE 7
Cross tabulation of the variable "Educational attainment" versus age group

Age-class	Educational attainment								
	Missing	0	1	2	3	4	5	6	9
0-4	47.674	0	0	0	0	0	0	0	898.187
5-9	1.414	0	0	0	0	0	0	0	1.009.745
10-14	1.275	0	0	0	0	0	0	0	977.689
15-19	27.192	147	258.459	539.275	163.161	43	16.684	0	0
20-24	216.918	1.830	14.448	126.841	405.943	2.987	209.128	4	0
25-29	476.523	3.800	10.723	36.161	147.537	2.779	312.380	79	0
30-34	589.228	4.661	11.692	22.976	104.633	3.041	288.873	217	0
35-39	830.625	5.136	12.811	27.478	112.666	5.832	280.059	399	0
40-44	996.017	4.888	12.890	31.923	100.996	7.906	148.114	485	0
45-49	1.022.110	4.596	13.128	33.829	79.051	8.578	91.443	502	0
50-54	955.668	4.245	14.695	32.938	58.110	7.787	68.153	409	0
55-59	944.786	4.018	16.745	30.682	41.449	6.790	45.224	392	0
60-64	881.054	3.536	16.656	28.929	30.744	5.720	28.315	335	0
65-69	656.135	2.962	11.408	19.367	18.166	3.353	13.914	190	0
70-74	547.283	2.059	8.049	12.352	10.552	1.818	6.867	101	0
75-79	457.713	1.254	5.361	8.994	6.520	1.094	3.900	51	0
80-84	324.613	493	3.356	6.650	3.922	617	2.321	35	0
85-89	181.535	201	1.911	3.531	1.811	262	909	13	0
90-94	64.749	85	852	888	496	61	268	7	0
95-99	14.174	22	201	157	78	10	40	2	0
100 or +	1.526	2	7	16	8	1	5	0	0

NOTE: Educational attainment: (0) No formal education, (1) ISCED level 1 Primary education, (2) ISCED level 2 Lower Secondary Education, (3) ISCED level 3 Upper Secondary Education, (4) ISCED level 4 Post Sec. non-tertiary study, (5) ISCED level 5 First stage of tertiary education, (6) ISCED level 6 Second stage of tertiary education, (9) Not applicable (persons < 15 yr).
Source: own calculations.

In table 8 the numbers of unemployed people, homemakers, and others come from the LFS meaning that for these categories only sample information is available. The results shown for these categories are not weighted to the population totals. Table 8 is in line with the fact that the pensionable age in the Netherlands is in general 65 years, i.e. there is a clear peak of records with a value of 3 (pension or capital income recipients) for the variable current activity status in the age groups 60-69. Related to this it can be seen that the part of people having status 1 (employed) significantly decreases once they have reached the age of 65 years. The status 4 (students not economically active) is also in line with the expectations as this status occurs mostly for people below the age of 25 years.

TABLE 8
Cross tabulation of the variable "Current activity status" versus age group

Ageclass	Current activity status							
	Missing	0	1	2	3	4	5	6
0-4	0	945.861	0	0	0	0	0	0
5-9	0	1.011.159	0	0	0	0	0	0
10-14	0	978.964	0	0	0	0	0	0
15-19	34.911	0	482.180	33	0	487.533	11	293
20-24	113.286	0	716.411	106	0	147.395	190	711
25-29	142.149	0	818.167	107	0	28.396	486	677
30-34	163.141	0	856.030	129	0	4.506	744	771
35-39	216.807	0	1.053.407	180	0	2.418	1.138	1.056
40-44	228.634	0	1.070.204	228	0	1.853	1.076	1.224
45-49	236.102	0	1.013.249	242	0	1.134	1.076	1.434
50-54	262.473	0	875.724	253	1	504	1.261	1.789
55-59	330.898	0	714.959	263	39.705	232	1.776	2.253
60-64	390.062	0	343.089	122	256.826	78	2.348	2.764
65-69	8.730	0	88.209	1	628.490	16	3	46
70-74	5.306	0	35.690	1	548.059	3	0	22
75-79	3.822	0	14.705	0	466.339	2	0	19
80-84	2.166	0	5.897	0	333.936	0	0	8
85-89	1.115	0	2.360	0	186.690	0	0	8
90-94	405	0	662	0	66.339	0	0	0
95-99	162	0	136	0	143.886	0	0	0
100 or +	97	0	18	0	1.450	0	0	0

NOTE: Current activity status: (0). Persons below minimum age for economic activity, (1) Employed, (2) Unemployed, (3) Pension or capital income recipients, (4) Students not economically active, (5) Homemakers, (6) Others.
Source: own calculations.

Based on these tables it can very cautiously be concluded that the values for the variables under concern are accurate. Although much more research on this topic is needed of course.

4. CONCLUSIONS

The virtual census has proved to be a successful concept in the Netherlands. It has many advantages compared to traditional censuses. The costs are now considerably lower. Still, Census data on the Netherlands can be compared to results of earlier Dutch censuses and to the results of other countries that take part in the same Census Round. So far the Netherlands has conducted three virtual censuses. However, the Dutch data that have been compiled for 1981 and 1991 were of a much more limited character than the set of tables of the 2001 Census. Moreover, they were largely based on a register count of the population in combination with the existing surveys on the labour force and housing conditions. Also for the Virtual Census of 2011 it is important that the final results are comparable both over time and with other countries. Therefore, the quality of the Dutch registers used is of vital importance for the 2011 Census.

It is possible to conduct a register-based census in more and more countries. Although in most countries, not all census variables can be derived from register information. For those variables additional surveys remain a necessity. To be able to use registers for statistical purposes, it should be possible to determine the quality of these registers. The results described in this paper show that the quality framework developed for administrative registers and the corresponding checklist are valuable tools for the evaluation of the statistical usability of such data sources. In 2012 it has been decided how the different Dutch Census variables had to be derived. In this decision making process the indicators in the Data hyper dimension played an important role. The variable current activity status is now a register variable, but no distinction is made between the categories homemakers and others. The variable educational attainment is based on the Labour Force Survey only as for the time being the value added of other sources is selective and adding 'half a register' would lead to difficult estimation problems.

An advantage of the approach used for the construction of the Virtual Census file (Schulte Nordholt, 2004) is the use of micro-integration. In this way data are checked and incorrect data are adapted. The number of measurement errors thus decreases. By the introduction of the technique of repeated weighting the remaining inconsistencies are solved. Given the detailed information requests of the 2011 Census, the available sources for the Dutch Census and our first experiences with applying the quality framework, it is sure that we will have a lot of interesting experiences with our register-based 2011 Census in the coming years that will draw the attention of many other countries.

REFERENCES

DAAS, P.J.H., OSSEN, S.J.L., VIS-VISSCHERS, R.J.W.M., and ARENDS-TOTH, J. (2009): "Checklist for the Quality evaluation of Administrative Data Sources", Discussion paper 09042 (Statistics Netherlands).

DAAS, P.J.H., OSSEN, S.J.L., and TENNEKES, M. (2010): The determination of administrative data quality: recent results and new developments, European Conference on Quality in Official Statistics 2010, Helsinki, Finland.

DAAS, P., OSSEN, S., TENNEKES, M., ZHANG, L-C., HENDRIKS, C., FOLDAL HAUGEN, K., CERRONI, F., DI BELLA, G., LAITILA, T., WALLGREN, A., and WALLGREN, B. (2011): "Report on methods preferred for the quality indicators of administrative data sources", Second deliverable of workpackage 4 of the BLUE Enterprise and Trade Statistics Project, September 28.

KARR, A. F., SANIL, A. P., and BANKS, D. L. (2006): "Data quality: A statistical perspective", Statistical Methodology, 3, 2, , pp. 137-173.

SCHULTE NORDHOLT, E. (2004): "Introduction to the Dutch Virtual Census of 2001", in SCHULTE NORDHOLT, E., HARTGERS, and M., GIRCOUR, R. (eds.), The Dutch Virtual Census of 2001, analysis and methodology, Voorburg, Statistics Netherlands, pp. 9-22.

UNITED NATIONS (2007): "Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics", United Nations, New York and Geneva.