Evaluation of the Quality of Administrative Data used in the Dutch Virtual Census*

Piet J.H. Daas[1], Saskia J.L. Ossen[1], Martijn Tennekes[1]
Eric Schulte Nordholt[2]
[1]Statistics Netherlands, Methodology sector, CBS-weg 11, 6412 EX, Heerlen, The Netherlands
[2]Statistics Netherlands, Division of Social and Spatial Statistics, Henri Faasdreef 312 2492 JP, The Hague, The Netherlands

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 administrative sources and surveys. This choice has lead to several methodological challenges. One of them is determining the effect of the quality of the sources on the combined result. For administrative sources this is a serious issue because the collection and maintenance is beyond the control of the Statistical Institute. It is therefore important that the Institute is able to determine the quality of such sources prior to use. For this purpose Statistics Netherlands has developed a quality framework. It consists of three high level views on the quality of administrative sources: a Source, a Metadata, and a Data view. The first two views are evaluated with a checklist that has already been applied successfully. Current research focuses on developing a systematic way to evaluate data quality. In this contribution the insights obtained in the research on the quality of administrative sources are applied to the Virtual Census.

Key Words: Quality, Registers, Administrative data, Checklist, Census.

1. Introduction

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 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 (Schulte Nordholt, 2012). 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.

* The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. This paper was presented at the Joint Statistical Meeting 2012, July 29-Aug. 2, San Diego, CA.

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; 2012). 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. 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.

2. 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*. In addition to register information, some information provided by the Labour Force Survey (LFS) remains essential for the 2011 Census.

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 variables. These are discussed below.

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 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.

3. 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., 2012). These views are referred to as hyperdimensions (Karr et al., 2006) and are called: Source, Metadata, and Data. Each hyperdimension 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. Subsection 3.1 starts with an overview of the quality aspects in the Source and Metadata hyperdimension and the methods developed to determine them. Next, recent insights on the study of the quality aspects in the Data hyperdimension are described.

3.1 Source and Metadata hyperdimensions

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 (Daas and Ossen, 2011). These components of quality are located in the Source hyperdimension of the quality framework. In table 1 the dimensions, quality indicators, and method descriptions for this hyperdimension are shown. The second hyperdimension in the framework, the Metadata hyperdimension, 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 hyperdimension.

For the evaluation of the quality indicators in the Source and Metadata hyperdimension 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 hyperdimensions. 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

Table 1: Quality framework for secondary data sources, Source hyperdimension.

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 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 data source is send to NSI
		Are security measures needed?(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 data source be delivered?
		Rate at which exceptions are reported
		Rate at which data is stored by DSH
	4.4 Format	Formats in which the data can be delivered
		Does this comply with the NSI-requirements?
5. Procedures	5.1 Data collection	Familiarity with the way the data is collected
	5.2 Planned changes	Familiarity with planned changes of 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
		Does this comply with NSI-requirements?

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

Table 2: Quality framework for secondary data sources, Metadata hyperdimension

Dimensions	*Quality indicators*	*Methods*
1. Clarity	1.1 Population unit definition	Clarity score of the definition
	1.2 Classification variable def.[3]	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 def. comp.[4]	Comparability with NSI definition
	2.2 Classification variable def. comp.	Comparability with NSI definition
	2.3 Count variable def. 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
	3.2 Unique combinations	Comparability with unique keys of NSI
		Presence of useful variable combinations
4. Data treatment (by DSH)	4.1 Checks	Population unit checks performed
		Variable checks performed
		Combinations of variables checked
	4.2 Modifications	Familiarity with data modifications
		Are modified values marked and how?

		Familiarity with default values used

[3] def.: definition; [4] comp.: comparison.

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 data in the file to the metadata
	1.3 Convertability	Conversion of the file to 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 value
	2.5 Inconsistent values	Extent of inconsistent values for combinations of variables
	2.6 Dubious values	Presence of implausible values or combinations of values
3. Completeness	*Objects*	
	3.1 Undercoverage	Absence of target objects in the source
	3.2 Overcoverage	Presence of non-target objects in the source
	3.3 Selectivity	Statistical coverage and representativity 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 actions by DSH
4. Time-related dimension	4.1 Timeliness	Time lag between the end of the reference period in the source and the moment of receipt
	4.2 Punctuality	Time lag between the settled data and actual delivery date
	4.3 Overall time lag	Time lag between the end of the reference period in the source and the moment NSI concluded that the data can be used
	4.4 Delay	Time lag between an actual change in the real-world and its registration in the source
	Objects	
	4.5 Dynamics	Changes in the population of objects over time
	Variables	
	4.6 Stability	Changes of variables or values over time
5. Integrability	*Objects*	
	5.1 Comparability of objects	Similarity of objects in source with objects used by NSI
	5.2 Alignment of objects	Linking-ability (align-ability) of objects in source with those of NSI
	Variables	
	5.3 Linking variable	Usefulness of linking variables (keys) in source
	5.4 Comparability of variables	Proximity (closeness) of variable values in different sources

mind, which is the 2011 Census in our case. The next step is the determination of the quality of the data (Daas et al., 2012).

3.2 Data hyperdimension

Indicators for the evaluation of the quality of the data in a register are part of the Data hyperdimension. 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., 2012). The indicators and dimensions identified are listed in table 3.

4. Quality evaluation results

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

4.1 Source and Metadata: application of the checklist

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

Table 4: Evaluation results for the Source hyperdimension

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 hyperdimension

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

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

4.2 Data: evaluation results

In this section indicators included in the data hyperdimension are discussed. In the available dataset, from hereon referred to as the Virtual Census test file, 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 *Integrability* are not considered here. Readers are referred to other papers of the authors for more detailed information on the dimensions and indicators used (Daas et al., 2011) and for examples (Daas et al., 2012). The analysis in this paper focuses on the *Accuracy* and the *Completeness* dimension with emphasis on the variables *level of education* (derived from the ER), and *current activity status* (derived from the UR, SR, and LFS).

4.2.1 Completeness dimension

One of the indicators in the completeness dimensions focuses on *selectivity*. This indicator looks at the statistical coverage and representativity of objects (units) in the data source. The latter can be illustrated by a visualization method specifically developed for the inspection of large data files; the so-called tableplot (Tennekes et al., 2011). The tableplot for the Virtual Census test file is shown in Figure 1. In this figure a selection of eight variables are displayed for a total of 16.5 million records (all registered Dutch inhabitants in 2008). Age is used as a sorting variable. Each column represents a variable and each row (‘bar’) is an aggregate of a fixed number of records (here a percentile). The numeric sorting variable ‘age’ is displayed as a bar chart (in blue) and the other variables are categorized and shown as stacked bar charts with a different color for each category.

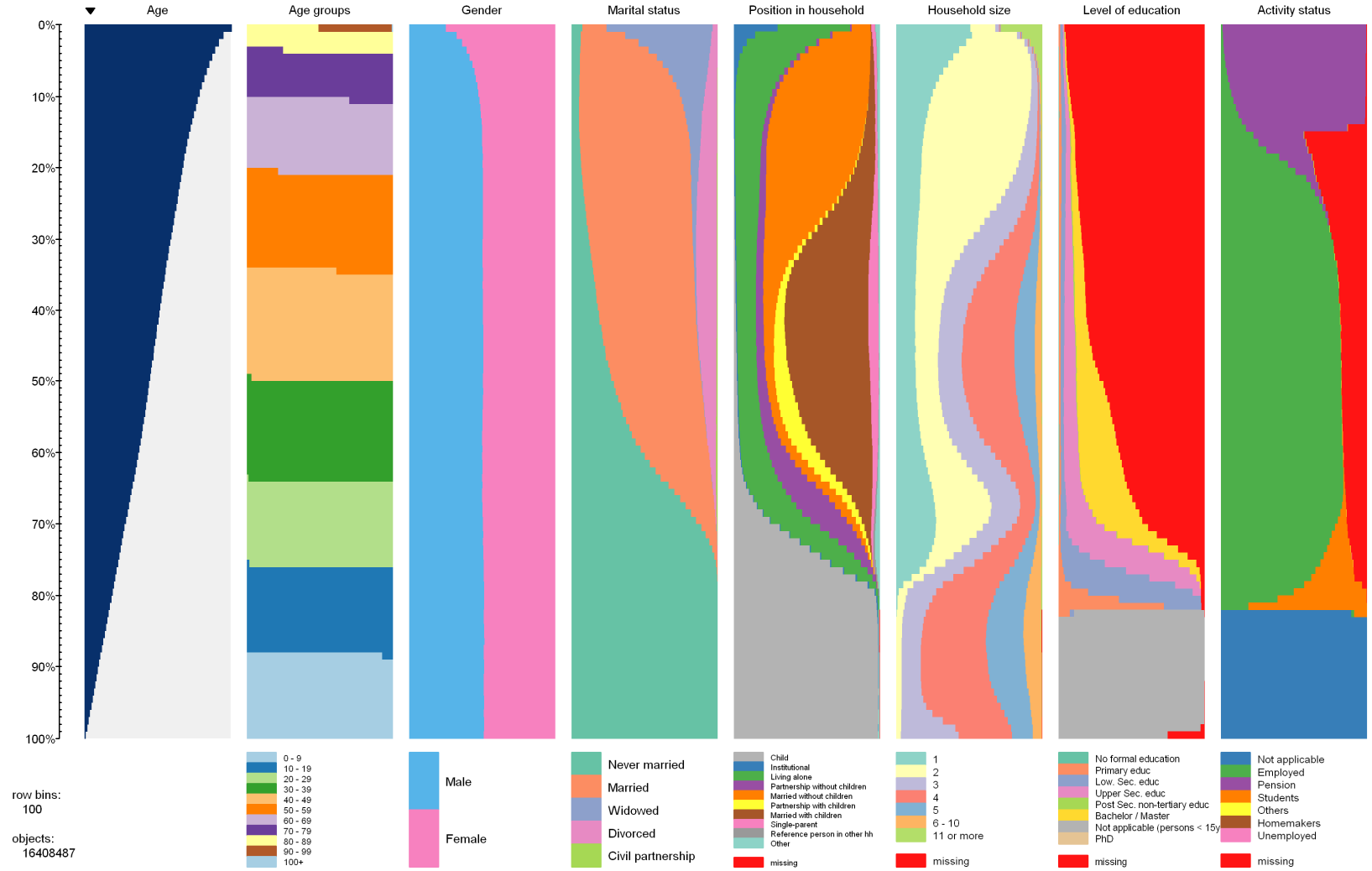

Figure 1: Tableplot of the 2008 Dutch Virtual Census test file. Age is used as the sorting variable (from high to low). The level of education and current activity status are shown in the seventh and eighth column, respectively.

The seventh column in Figure 1 displays the various categories for the *level of education* and illustrates the occurrence and distribution of missing values; shown in red. In the Netherlands, people over 15 can have various levels of education. However, with increasing age the amount of missing information increases dramatically. This is caused by the fact that the official registration of the level of education of graduates has only recently started in the Netherlands. As a result, only for people that have recently finished school, the highest level of education is nearly completely available; these are predominantly young people. For all others, data in sample surveys and data available in other specific educational registrations are jointly used to derive the highest level of education obtained (Bakker et al., 2008). The latter information is only available for a limited part of the population which explains the increasing number of missing values with increasing age. Under 15, people do not have a formal level of education and should be categorized as 'not applicable'. The tableplot also shows that the lowest two rows in the seventh column, clearly contain a considerable number of missing values. This is an obvious error that needs to be corrected.

The eighth column reveals that data on *current activity status* are also selectively missing for people between 15 and 65 years. This 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 the dataset. What column eight also reveals is that a considerable number of elderly people apparently remain economically active in the Netherlands (more on this in the next section).

Another indicator regarding completeness is *redundancy*, i.e. the presence of multiple registrations of objects. To investigate whether or not the file suffered from redundancy, we searched for rows in our dataset which showed equal values for all variables (including *level of education* and *current activity status*) with exception of the unique persons identification number used. 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 institutes. 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. The reader is referred to Tennekes et al. (2013) for a more detailed description of the wealth of information provided by tableplots.

4.2.2 Accuracy dimension

Regarding the *accuracy* dimension we consider in this subsection whether there are any dubious values in the data. Here, we concentrate on the variable *current activity status*. Here again the relation with age is interesting. For example, it is expected that (almost) only elderly people will have a value for the *current activity status* equal to 3 (pension or capital income recipients). To check the relations between these variables in detail, cross tabulations were created. The results are shown in table 6.

In table 6 the numbers of unemployed people (column 2), homemakers (column 5), and others (column 6) 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 6 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

Table 6: Cross tabulation of the variable 'Current activity status' versus age group[5]

Ageclass	Current activity status							
	Missing	0	1	2	3	4	5	6
1: [0, 5)	0	945861	0	0	0	0	0	0
2: [5, 10)	0	1011159	0	0	0	0	0	0
3: [10, 15)	0	978964	0	0	0	0	0	0
4: [15, 20)	34911	0	482180	33	0	487533	11	293
5: [20, 25)	113286	0	716411	106	0	147395	190	711
6: [25, 30)	142149	0	818167	107	0	28396	486	677
7: [30, 35)	163141	0	856030	129	0	4506	744	771
8: [35, 40)	216807	0	1053407	180	0	2418	1138	1056
9: [40, 45)	228634	0	1070204	228	0	1853	1076	1224
10: [45, 50)	236102	0	1013249	242	0	1134	1076	1434
11: [50, 55)	262473	0	875724	253	1	504	1261	1789
12: [55, 60)	330898	0	714959	263	39705	232	1776	2253
13: [60, 65)	390062	0	343089	122	256826	78	2348	2764
14: [65, 70)	8730	0	88209	1	628490	16	3	46
15: [70, 75)	5306	0	35690	1	548059	3	0	22
16: [75, 80)	3822	0	14705	0	466339	2	0	19
17: [80, 85)	2166	0	5897	0	333936	0	0	8
18: [85, 90)	1115	0	2360	0	186690	0	0	8
19: [90, 95)	405	0	662	0	66339	0	0	0
20: [95, 100)	162	0	136	0	14386	0	0	0
21: [100, ∞)	97	0	18	0	1450	0	0	0

[5] 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.

groups 60-70. 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, although there seem to be quite some people above 70-75 years that are still working. The latter suggests an accuracy issue that requires further attention. The status 4 findings (students not economically active) are in line with the expectations as this status occurs mostly for people below the age of 25 years. All these findings are also indicated in the Tableplot in figure 1, but the information displayed in table 6 is -obviously- more quantitative as it displays exact numbers.

Based on the table and tableplot figure it can be -very cautiously- concluded that the majority of the values for the variables under concern seem accurate. Although much more research on this topic is needed of course. Especially the employed people above 70-75 included in the Virtual Census test file require further attention.

5. 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 and census data on the Netherlands can still be compared to results of earlier Dutch censuses and those 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 then existing LFS and survey on 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 (Schulte Nordholt, 2012). Therefore, the quality of the Dutch registers used is of vital importance for the 2011 Census.

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. Since the main Census project continues until 2014, it will be decided in the coming years how the different Dutch Census variables will be derived. During that period more of the indicators in the Data hyperdimension will be applied to the data in the registers used.

Big advantage of the approach used for the construction of the Virtual Census file (Schulte Nordholt, 2004; 2012) 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

Bakker B.F.M., Linder, F., van Roon, D. (2008), Could that be true? Methodological issues when deriving educational attainment from different administrative datasources and surveys. Paper for the IAOS Conference on Reshaping Official Statistics, October 14-16, Shanghai, China.

Daas, P.J.H., Ossen, S.J.L. (2011) Metadata Quality Evaluation of Secondary Data Sources. *International Journal for Quality Research*, 5 (2), 57-66.

Daas, P.J.H., Ossen, S.J.L., Vis-Visschers, R.J.W.M., 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., Tennekes, M., Burger, J. (2012) Evaluation and visualisation of the quality of administrative sources used for statistics. Paper for the European Conference on Quality in Official Statistics 2012, May 29-June 1, Athens, Greece.

Daas, P., Ossen, S., Tennekes, M., Zhang, L-C., Hendriks, C., Foldal Haugen, K., Cerroni, F., Di Bella, G., Laitila, T., Wallgren, A., 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., Banks, D. L. (2006) Data quality: A statistical perspective. *Statistical Methodology*, 3, 137-173.

Schulte Nordholt, E. (2004) Introduction to the Dutch virtual census of 2001. In: *The Dutch Virtual Census of 2001, analysis and methodology*, Eds., Schulte Nordholt, E., Hartgers, M., Gircour, R., Statistics Netherlands, Voorburg, pp. 9-22.

Schulte Nordholt, E. (2012) Methodology used for estimating Census tables based on incomplete information. Paper for the UNECE-Eurostat Expert Group Meeting on Censuses Using Registers, May 22-23, Geneva, Switzerland.

Tennekes, M., de Jonge, E., Daas, P.J.H. (2011), Visual Profiling of Large Statistical Datasets. Paper for the New Techniques and Technologies for Statistics conference, February 22-24, Brussels, Belgium.

Tennekes, M., de Jonge, E., Daas, P.J.H. (2013) Visualizing and Inspecting Large Datasets with Tableplots. *Journal of Data Science* 11(1), in press.