Distr.: General
27 March 2012

Original: English

Economic Commission for Europe

Conference of European Statisticians

UNECE-Eurostat Expert Group Meeting on Censuses Using Registers

Geneva, 22-23 May 2012
session 3: Availability, completeness and quality of data from registers and other sources

Quality of registers used for the Dutch census

Note by Statistics Netherlands[1]

Summary

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.

Independently from the research on improving the Virtual Census, Statistics Netherlands is busily involved in developing systematic methods for determining the quality of registers. For this purpose a quality framework has been developed. The framework consists of three high level views on the quality of administrative data: the Source view, the Metadata view, and the Data view. To evaluate the first two views in a systematic way a checklist has been developed which has already been successfully applied to several data sources. Current research focuses on developing a systematic way for evaluating data quality.

In this contribution we explore the possibilities for using the insights obtained in the

[1] Prepared by Eric Schulte Nordholt, Saskia Ossen and Piet Daas.

.

research on the quality of registers to improve decisions made in the Virtual Census about which sources should be used for deriving or estimating which variables. This is also beneficial for the research on quality as it can be tested whether the approach developed is also applicable to censuses.

I. Introduction

1. 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. 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.

2. 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. 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.

II. Data sources and variables

3. 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.

4. 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.

5. 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.

6. *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?

7. *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.

8. 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.

III. Quality framework

9. 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).

10. 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.

11. 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.

12. 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, ranging from good to poor, is stated. 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.

13. 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 is on quality of the data in the registers used as input in the statistical process. The quality dimensions identified for data are Technical checks, Accuracy, Completeness, Time-related dimension and Integrability (Daas et al., 2011).

IV. Quality evaluation results

14. 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.

A. Source and Metadata: application of checklist

15. 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 1 and 2, respectively.

Table 1: 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 2: 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	+	+	+	+	+

15. 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.

16. The results in table 1 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.

17. The results in table 2 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.

18. 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.

B. Data: first evaluation results

19. 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: 1 January 2008. This implies that the indicators referring to the dimensions: *Technical checks*, *Time-related*, and *Integrability* are not considered in this paper. The analysis therefore rather focuses on the *Completeness* and the *Accuracy* dimension.

1. Completeness dimension

20. 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 will become available at a later stage and are therefore not discussed further in this paper.

21. 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 3 provides an overview of the number of missing values for both variables studied.

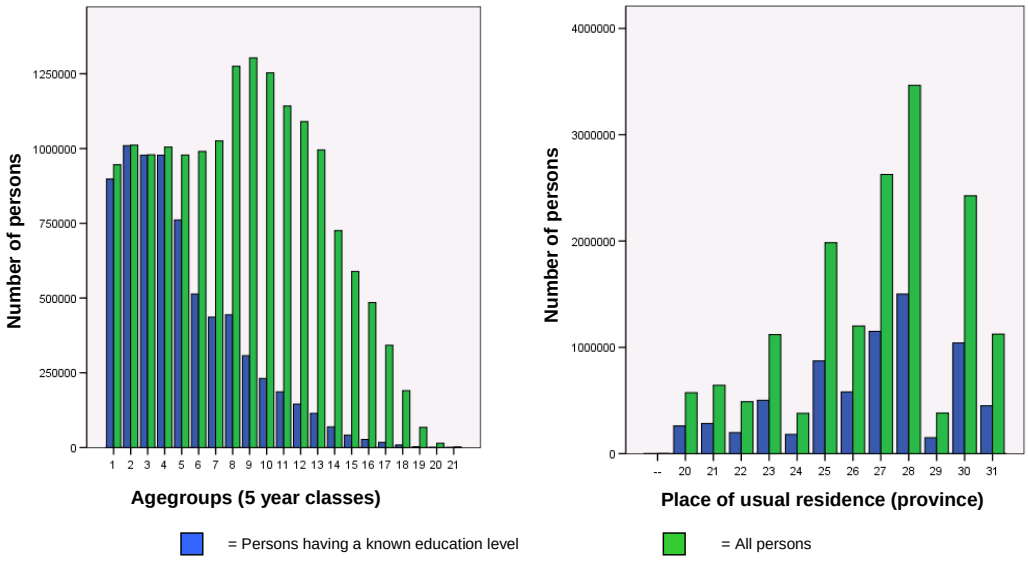
Table 3: Overview of missing values

Variable	*Number of missings*	*Percentage missing (%)*
Educational attainment	9,238,212	56.3
Current activity status	2,140,266	13.0

22. 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 young people. 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).

23. 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.



24. 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 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.

2. Accuracy dimension

25. 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).

26. 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 4 and table 5. 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 4). 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 20-25. Furthermore, there turn out to be some people who reached educational level 5 (First stage of tertiary education) already within the age group 15-20. 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 4: Cross tabulation of the variable "Educational attainment" versus age group

Ageclass	Educational Attainment								
	Missing	0	1	2	3	4	5	6	9
1: [0, 5)	47674	0	0	0	0	0	0	0	898187
2: [5, 10)	1414	0	0	0	0	0	0	0	1009745
3: [10, 15)	1275	0	0	0	0	0	0	0	977689
4: [15, 20)	27192	147	258459	539275	163161	43	16684	0	0
5: [20, 25)	216918	1830	14448	126841	405943	2987	209128	4	0
6: [25, 30)	476523	3800	10723	36161	147537	2779	312380	79	0
7: [30, 35)	589228	4661	11692	22976	104633	3041	288873	217	0
8: [35, 40)	830625	5136	12811	27478	112666	5832	280059	399	0
9: [40, 45)	996017	4888	12890	31923	100996	7906	148114	485	0
10: [45, 50)	1022110	4596	13128	33829	79051	8578	91443	502	0
11: [50, 55)	955668	4245	14695	32938	58110	7787	68153	409	0
12: [55, 60)	944786	4018	16745	30682	41449	6790	45224	392	0
13: [60, 65)	881054	3536	16656	28929	30744	5720	28315	335	0
14: [65, 70)	656135	2962	11408	19367	18166	3353	13914	190	0
15: [70, 75)	547283	2059	8049	12352	10552	1818	6867	101	0
16: [75, 80)	457713	1254	5361	8994	6520	1094	3900	51	0
17: [80, 85)	324613	493	3356	6650	3922	617	2321	35	0
18: [85, 90)	181535	201	1911	3531	1811	262	909	13	0
19: [90, 95)	64749	85	852	888	496	61	268	7	0
20: [95, 100)	14174	22	201	157	78	10	40	2	0
21: [100, ∞)	1526	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.

Table 5: Cross tabulation of the variable "Current activity status" versus age group

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 5
9: [40, 45)	228634	0	1070204	228	0	1853	1076	1224 4
10: [45, 50)	236102	0	1013249	242	0	1134	1076	1434 4
11: [50, 55)	262473	0	875724	253	1	504	1261	1789 9
12: [55, 60)	330898	0	714959	263	39705	232	1776	2253 3
13: [60, 65)	390062	0	343089	122	256826	78	2348	2764 4
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

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.

27. In table 5 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 5 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, although there seem to be some people older than 100 years that are still working. 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.

28. 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.

V. Conclusions

29. 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 then 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.

30. 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 the coming years it will be decided how the different Dutch Census variables will be derived. During that period more of the indicators in the Data hyper dimension will be applied to the data in the registers used.

31. Big 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., S.J.L. Ossen, R.J.W.M. Vis-Visschers and J. Arends-Toth, 2009. Checklist for the Quality evaluation of Administrative Data Sources. Discussion paper 09042, Statistics Netherlands, The Hague / Heerlen. http://www.cbs.nl/NR/rdonlyres/0DBC2574-CDAE-4A6D-A68A-88458CF05FB2/0/200942x10pub.pdf

Daas, P.J.H., S.J.L Ossen and M. Tennekes, 2010. The determination of administrative data quality: recent results and new developments. Paper for the European Conference on Quality in Official Statistics 2010, Helsinki, Finland.

Daas, P., S. Ossen, M. Tennekes, L.C. Zhang, C. Hendriks, K. Foldal Haugen, F. Cerroni, G. Di Bella, T. Laitila, A. Wallgren and B. Wallgren, 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, 28 September 2011.

Karr, A.F., A.P. Sanil and D.L. Banks, 2006. Data quality: A statistical perspective. *Statistical Methodology*, Number 3, pp. 137-173.

Schulte Nordholt, E., 2004. Introduction to the Dutch Virtual Census of 2001. *The Dutch Virtual Census of 2001, analysis and methodology*, eds. E. Schulte Nordholt, M. Hartgers and R. Gircour, Statistics Netherlands, Voorburg, pp. 9-22.