Linking Register of Construction Works with Census in Estonia

Kai Kaarna

Statistics Estonia, Estonia
e-mail: kai.kaarna@stat.ee

Abstract

In 2007 in Estonia were carried out the project "Preparation for the 2011 Census: Quality evaluation of the Register of Construction Works". The basis for evaluation was a random sample of 4,700 buildings from the 2000 Population and Housing Census database.
There was no unique identifier for buildings in both databases and available address characteristics were used for linking.

1 Introduction

The aim of this paper is to describe linking process and give some examples of data quality in Register of Construction Works (RCW).

In autumn 2005 Statistics Estonia (SE) started to plan the next Population and Housing Census, which will be conducted in 2011.

For independent evaluation of the quality and usage of the Register of Construction Works data in statistics, Statistics Estonia together with the Ministry of Economic Affairs and Communications (MEAC) carried out the project "Preparation for the 2011 Census: Quality evaluation of the Register of Construction Works".

The activities concerning that project were connected with the linking of two different databases, comparing the data in them and mapping activities which are necessary to convert to register-based capitation.

For record linking there were no unique identifiers, but there was possibility to link buildings and dwellings from the last Census and from RCW by addresses. Addresses were divided into following components: state, county, town or rural municipality, settlement, street or farm name, number of building and number of dwelling. We had had to link records by these available variables. But that could be affected by errors.

The quality analysis of characteristics provided information about the quality of the characteristics necessary for the production of Census statistics.

2 Design of sample

The basis for linking was a random sample of 4,700 buildings from the 2000 Population and Housing Census database. The stratas had been formed considering the location of the building (county and settlement type). All the buildings were divided into 47 layers, from each 100 buildings were randomly selected to be used in the survey.

Stratas were defined by settlement type in the following way:

- Cities (number of inhabitants ≥ 50 000 (6 stratas);
- Other county centrums (11 stratas);
- Towns and hamlets (15 stratas);
- Villages (15 stratas)

Of the buildings in the random sampling, the sample of dwellings was taken as follows: from the buildings with up to 4 dwellings all dwellings were taken into the sample and from the rest of buildings every tenth dwelling. The sample of dwellings included 6,193 dwellings.

In Estonian villages there have been historically used both farm names and street names. In the beginning of 20 century there were only farm names and every farm was identifiable. Most of Estonian villages are scattered and there are no street names and numbers of the buildings. In Soviet time farms were reorganized and historical farm names were not used and have been forgotten and don't correspond to the present houses. Because of historical changes, there are some buildings addressed by farm names or lately by street names with number of buildings but some are in the register only by name of village.

3 Linking process

In the linking process our attempt was to find for all items of the sample a “partner” from RCW if possible.

There were three stages for linking buildings. At first an attempt was made to locate an automatic response to each building from the sample amongst the RCW, thereafter the buildings not linked were checked manually, one by one, and if possible were linked. At the same time the reasons for unlinking were studied and new regulations were made for automatic linking. Then a new automatic linking attempt was made.

All the (sample) dwellings of the linked buildings were used for linking dwellings.

3.1 1:1 or not

In some cases several “partners” were found to an address from Census. There were many records belonging to the register, which matched by address with same record from the census. Obviously sometimes among these buildings were cotes, sheds etc. We checked if there was living space in the partner and then matched the Census-address with that record. We decided that there is unique match (linked uniquely) if in RCW can be found only one linkable building with living space.

3.2 Manual linking

While linking the buildings manually we were discovered by regions many different types of errors that caused unlinking. The most commonly the reason was in writing stile of the texts in addresses. Hence abbreviations, but also quotation marks, first name expansion, special letters, dash and space differences in street/farm names were the reasons for unlinking. These reasons were taken into consideration for generating new rules for linking buildings by addresses.

3.3 New automatic linking

It was decided to apply some new rules and to carry out the third linking. The rules applied well and as a result of the third linking we succeeded to link, within the whole sampling, 68% (3,188) of buildings.

3.4 Some results of linking buildings

49% of the buildings (2,304 buildings) were linked using the first program, uniquely only 69% of them.

In manual linking stage there were linked uniquely 444 buildings (14% of uniquely linked buildings) and for 540 (41% of unlinked buildings) it was recognized that linking is not possible.

In additional automatic linking stage there was linked 68% of the buildings sample and uniquely 1,977 buildings (95% of uniquely linked buildings).

Table 1: *Building linking by stages*

	I automatic		manual		II automatic		Sample, N
Result	N	%	N	%	N	%	
Equivalent is a building with dwelling(s)	2,184	69	444	14	2,977	95	3,146
Equivalents are several buildings with dwelling(s)	104	56	21	11	173	93	187
Equivalent is buildings with no dwelling(s)	16	33	17	35	38	78	49
Not linked	0	0	540	41	0	0	1,318
Sample	2,304	49	1,022	22	3,188	68	4,700

3.5 Uniform address-standard

In Estonia standardized address data system (ADS) has been developed now but it was not used in these databases. Since up to the summer 2007 the uniform address-standard has been absent in the country. ADS could solve these linking problems, but only if it will be used in both (register and Census data).

3.6 Some more results of linking buildings and dwellings

Using the expansion factors (weights) we expanded the results for the whole Census database.

Weighted results are not so good: we can uniquely link 55% of buildings and in some counties even less than 25%. In one strata (villages in county of Võru) there were no linked buildings at all. This area is a periphery with especially little villages.

We were able to link at least buildings for 72% of dwellings totally. In some of these cases only one part of the address – the number of dwelling – did not match in both data-bases. In county of Ida-Viru we linked buildings for 85% of dwellings and in county of Viljandi for 84% but in county of Põlva only for 32% and in county of Võru for 40%.

Table 2: *Building linking by counties (on the assumption of buildings), weighted*

County	Equivalent(s)				sample	Total
	one building with dwelling(s)	several buildings with dwelling(s)	buildings with no dwelling(s)	Not linked		
	%	%	%	%	N	N
Harju	68.5	7.0	1.7	22.8	600	39,355
Hiiu	71.6	2.7	1.5	24.2	200	3,328
Ida-Viru	60.2	3.4	1.3	35.0	300	13,289
Jõgeva	27.5	0.5	0.2	71.7	300	10,210
Järva	50.0	1.5	-	48.5	300	8,910
Lääne	60.4	2.4	0.4	36.9	300	7,309
Lääne-Viru	42.0	1.9	0.9	55.3	300	14,863
Põlva	15.9	0.6	0.3	83.2	300	9,700
Pärnu	61.3	3.5	1.5	33.7	300	17,788
Rapla	70.0	2.8	1.9	25.4	300	9,544
Saare	66.9	4.9	1.6	26.5	300	10,312
Tartu	42.3	4.1	0.4	53.2	300	21,096
Valga	60.7	3.9	0.2	35.1	300	8,637
Viljandi	82.7	4.7	2.8	9.8	300	13,207
Võru	21.2	2.2	0.3	76.3	300	10,146
Total	54.9	3.8	1.1	40.2	4,700	197,694

Table 3: *Buildings linking by counties (on the assumption of dwellings), weighted*

County	Equivalent(s)				Sample (dwellings)	Total (dwellings)
	one building with dwelling(s) %	several buildings with dwelling(s) %	buildings with no dwelling(s) %	Not linked %	N	N
Harju	78.9	10.0	0.8	10.3	967	224,763
Hiiu	74.7	2.7	1.3	21.4	224	5,003
Ida-Viru	85.0	4.2	2.0	8.8	539	85,859
Jõgeva	42.0	0.6	0.3	57.1	334	17,951
Järva	63.6	2.1	-	34.3	374	18,558
Lääne	69.9	2.2	0.4	27.5	365	15,145
Lääne-Viru	57.5	3.4	0.7	38.4	384	33,256
Põlva	31.8	0.9	0.5	66.8	384	15,656
Pärnu	72.7	4.4	1.0	21.9	383	40,127
Rapla	72.0	4.7	2.5	20.7	365	17,551
Saare	72.2	5.8	1.2	20.8	338	16,453
Tartu	63.0	7.5	0.4	29.2	408	64,660
Valga	70.4	3.9	0.2	25.5	366	17,440
Viljandi	83.6	4.5	2.2	9.7	377	25,954
Võru	39.8	6.8	0.4	53.0	385	18,891
Total	71.9	6.4	1.0	20.7	6193	617,267

The equivalents could be found for 43-44% of dwellings (in some cases in register there were no parts as dwellings but there have been made changes in the register) and for 28% of dwellings at least buildings could be linked although the dwellings could not be linked.

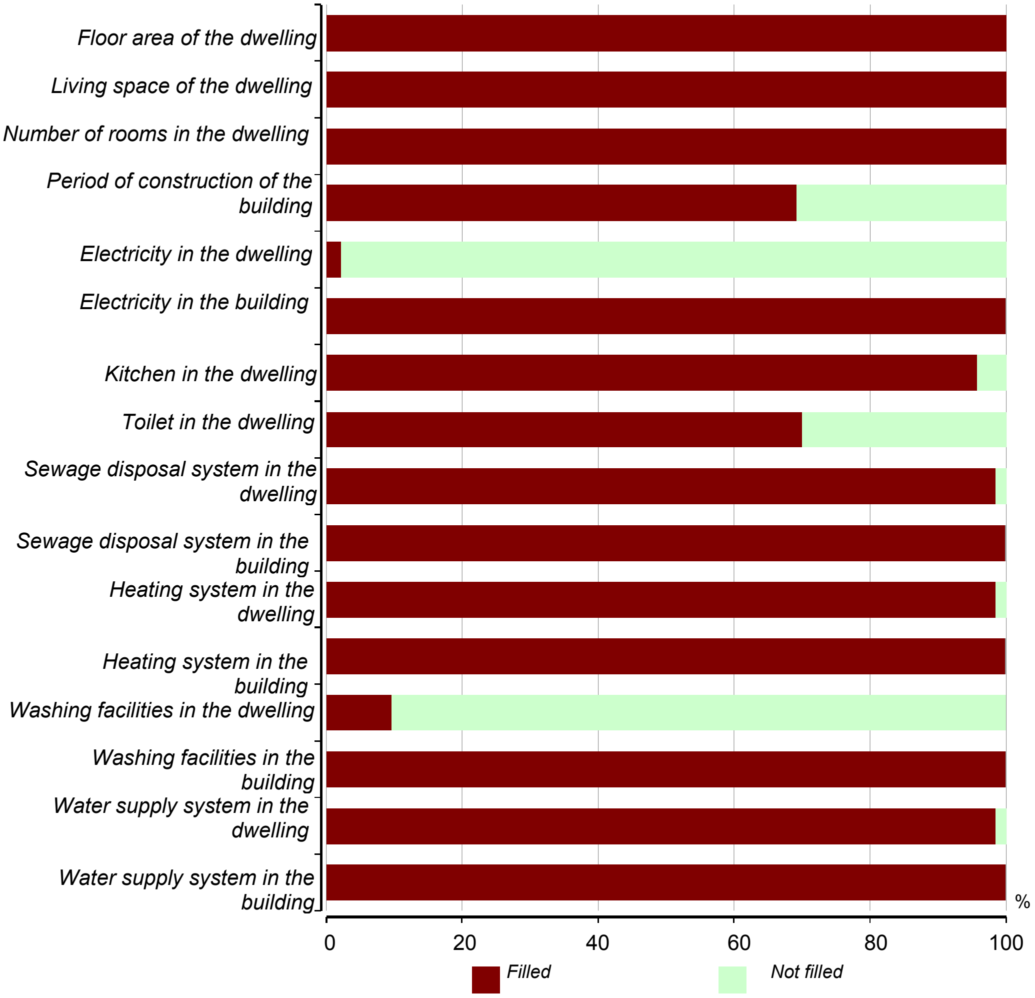
Table 4: *Buildings and dwellings linking, weighted*

Buildings	Dwellings	Dwellings, %
Equivalent is a building with dwelling(s)	Linked with RCW part of building	43.4
	Linked with RCW building (building is not divided in parts)	0.2
	Dwelling is not linked, but RCW is divided in parts	28.3
Equivalent is several buildings with dwelling(s)	Also the building is unlinked	6.4
Equivalent is buildings with no dwelling(s)	Also the building is unlinked	1.0
Not linked	Also the building is unlinked	20.7
Total (N)		617,267
Sample (N)		6,193

4 The quality of characteristics in RCW

Although the Register of Construction Works (RCW) comprises the most important variables for Census, the fulfillment of the variables and quality of data is not sufficient for producing reliable Census statistics.

Figure 2 *Completeness of the characteristics in the Register of Construction Works*



4.1 Number of dwellings in the building

Number of dwellings is given for all buildings in both data sources. For 85% the numbers are exactly the same and for 10% the numbers of dwellings differ by one.

Table 5: *Differences in number of dwellings of building, weighted*

Number of dwellings in the building	From dwellings. weighted
1 more by Census	5.5%
1 less by Census	4.6%
2 more	0.7%
2 less	1.4%
more than 2 more	0.1%
more than 2 less	2.8%
same	84.9%
Total	100.0%

4.2 Year of construction

RCW indicates the time of construction as a year, the Census — as a period. The year of construction is not filled for 31% of the sample.

In RCW, in case of 31% of dwellings the year of construction is missing, but in the Census2000 the period of the construction of the corresponding building has been recorded. 7% of buildings had no construction year according to both data sources.

In order to enable comparison, the construction time of RCW was divided into periods and the result was that 46% of the linked dwellings were in the comparable periods; in case of 11% the year of construction was different.

While Census data are from year 2000 and RCW data from year 2007, we decided that the year of construction is suitable, when according to the Census it was 1996 or later or the building was not finished and according to the RCW the year of construction was 1996–2007.

Tabel 6: *Comparison of year of construction, weighted*

	Number of dwellings in a building		
Year of construction	Different	Same	All
Same	25.1%	46.6%	44.0%
Anticipated	0.0%	2.3%	2.1%
Unknown in both	17.3%	5.9%	7.3%
Unknown in Census, known in RCW	6.8%	5.0%	5.2%
Known in Census, unknown in RCW	40.8%	29.2%	30.5%
Different	10.0%	11.0%	10.9%
Total	100.0%	100.0%	100.0%

4.3 Electricity

In everyday life most of Estonians are used to having electricity. In RCW the characteristic of electricity has been completed for only 1% of dwellings.

As at dwelling's level the information about electricity is missing in RCW in 99% of cases, in the analysis the value of electricity characteristic of the building was then taken into account. The outcome was logical. 96% of electricity characteristics had the same value according to RCW and the Census, 1.3% had different values and only in 2.4% of cases the value was known according to the Census, but unknown according RCW. In 0.1% of cases the value was known according to RCW, but was unknown according to the Census. Obviously it is not considered important to fill the electricity information at dwelling level, as well as at building level, but the characteristic is available for both levels.

Table 7: *Electricity in dwelling, weighted*

	Type of building and number of dwellings of building			
Electricity	Different	Same	Other	From dwellings
Same (is/is not)	1,2%	1,3%	7,9%	2,3%
Unknown in both	0,0%	0,2%	0,0%	0,1%
Known in Census, unknown in RCW	98,8%	98,5%	92,1%	97,5%
Different	0,0%	0,0%	0,0%	0,0%
Total	100,0%	100,0%	100,0%	100,0%

Table 8 *Electricity (changed)* [1]*, weighted*

	Type of building and number of dwellings of building			
Electricity	Different	Same	Other	From dwellings
Same (is/is not)	94,3%	96,3%	95,9%	96,2%
Known in Census, unknown in RCW	5,3%	2,2%	2,9%	2,4%
Unknown in Census, known in RCW	0,0%	0,2%	0,0%	0,1%
Different	0,4%	1,3%	1,2%	1,3%
Total	100,0%	100,0%	100,0%	100,0%

References

TF2005 project “Preparations for the 2011 Population Census: evaluation of the quality of the Register of Construction Works” in Estonia, *Final Report* (2008).

[1] If at dwelling’s level the information about electricity was missing in RCW then the information of the building was taken into account.