Geocoding of Population and Housing Census 2000 Data

Dāvis Kļaviņš, GIS specialist | davis.klavins@csb.gov.lv

12.01.2018

Research on topics like migration and depopulation in space and time requires longitudinal geocoded dataset on population, employment and housing conditions. The most complete respective data can be obtained only within population censuses.

Unlike data of the Population and Housing Census (PHC) 2011, the data of PHC 2000 have not been geocoded so far, therefore their use in longitudinal research was limited.

Within the framework of the Eurostat grant project:

- **geocoding of PHC 2000 dataset was made**; and

- population indicators were published, including in 1×1 km grid for whole Latvia and in 100×100 m grid for cities under state jurisdiction, moreover in remotely accessible anonymized dataset for scientific use (similarly as PHC 2011 dataset is published).

Information on addresses from questionnaires was recognized optically, and it contains errors:

House No.	Street name
/1	TARTU
11	
11	TART4
11	TARTU
2A	TA.RTU .
2A	TAR+4
2A	TART4
2A	TARTU
2A	TART-U
-2A	TARTU
5	
5	TART4
5	TARTU
5	TARTU -
5	TARTU.
ZA	TARTU

By linking information on address in PHC 2000 and in Population Register (PR) as of 1 March 2000, the probability that addresses match was determined.

Address data in PR still contain format errors, but much less than in PHC 2000 dataset.

Probabilities were calculated in R software, by using fields that form an address (street/house name, house No. etc.).

For each field of all the records, the Levenshtein distance (minimum number of changes required to transform one word into the other) was calculated. Afterwards, it was divided by length of the longest of both words and subtracted from 1, thereby calculating the probability that the fields are equal.

DAUGAVPIL.5 → DAUGAVPILS (distance is equal to 2)

$$\text{Probability} = 1 - \frac{2}{11} \approx 0.82$$

There are special cases when one word fits completely into another. Even though fields often describe the same geographic feature, the probabilities calculated were low, since one of the fields was shortened. Therefore probability calculation for such cases were adjusted:

KR BARONA → KRIŠJĀŅA BARONA

$\text{Probability}_{\text{before}} = 0.6$

$\text{Probability}_{\text{after}} \approx 0.87$

To calculate the combined probability for each record, the probabilities were combined with weights.

By using MS SQL Server and based on the form the address is written, addresses from PHC 2000 dataset in several groups (~30) were linked with
1) historical data from the State Address Register (SAR); or if linking was not possible,
2) PHC 2011 dataset.

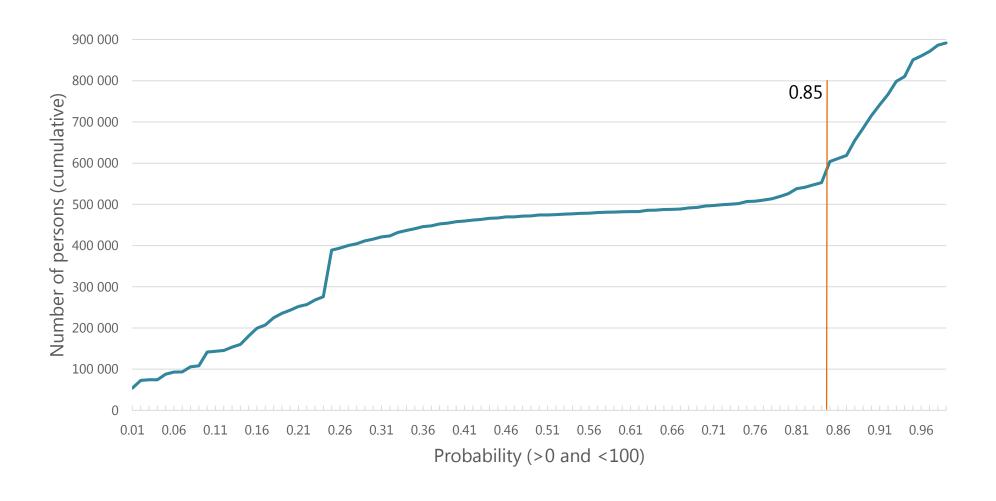Linked datasets contain coordinates to be added to the PHC 2000 dataset.

In some groups, calculated probabilities were used to substitute address information from questionnaires with that in the PR.

E.g., if probability of address equality between the information from questionnaires and PR is >0.85, the address from PR was used for linking with the SAR (linked fields – village name, street name and house No.).

Probability that address in questionnaires and PR is the same

For the remaining addresses, probability of their equality between PHC 2000 and PHC 2011 datasets was determined and under several conditions used to assume that the address from the PHC 2011 can be used.

E.g., if probability of equality between house name in the questionnaire in the PHC 2000 dataset and village name in the PHC 2011 dataset is ≥0.7 (in the PHC 2000 dataset, only village name was given in the field of house name, and address from PHC 2011 can be used if according to PHC 2011 at least one person from the household lives at the same village).

For addresses which could not be linked automatically and cover ~245 000 persons (10%), probabilities of address equality between PHC 2000 and SAR were determined. The 10 highest probabilities for each address were used for manual linking (all if the highest probability was equal for more than 10 addresses).
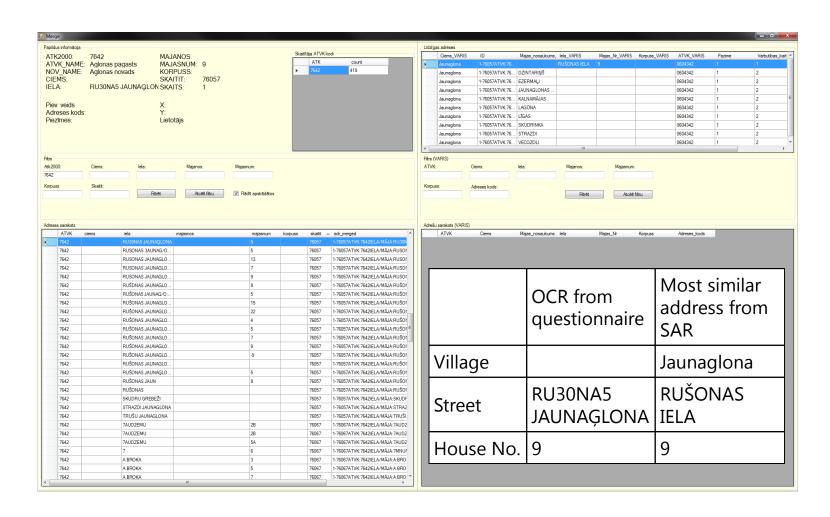
The number of records to be linked manually is considerably smaller than ~245 000 because of the grouping by address (initially those were ~37 000, but in some cases addresses had to be split into households or persons).
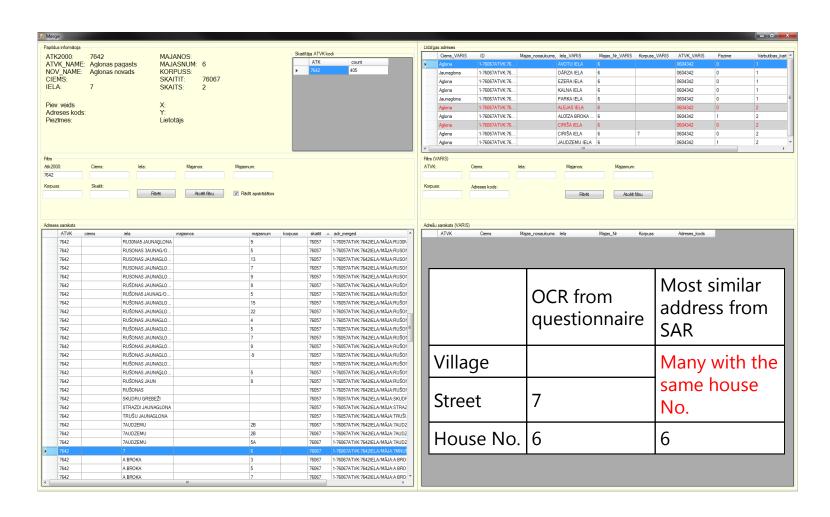
With an aim to facilitate manual linking, a software was developed, in which apart from the SAR addresses with the highest probabilities of being equal also access to the following sources was given:

- scanned questionnaires;

- all addresses surveyed by the interviewer;

- registered places of residence (until 2016) and address in the PHC 2011.
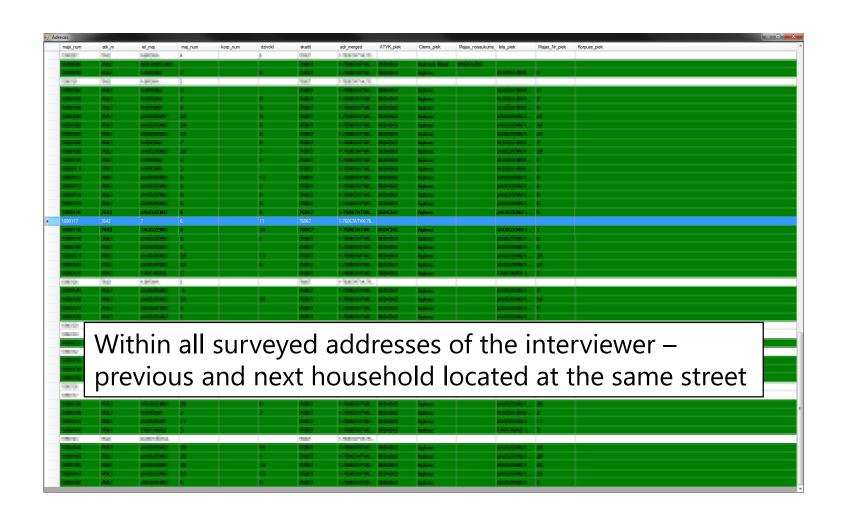
Work with the software for manual linking of addresses

	OCR from questionnaire	Most similar address from SAR
Village		Jaunaglona
Street	RU30NA5 JAUNAĢLONA	RUŠONAS IELA
House No.	9	9

Work with the software for manual linking of addresses

	OCR from questionnaire	Most similar address from SAR
Village		Many with the same house No.
Street	7	
House No.	6	6

Work with the software for manual linking of addresses

Within all surveyed addresses of the interviewer – previous and next household located at the same street

Work with the software for manual linking of addresses

Search for the correct address in SAR and link it manually
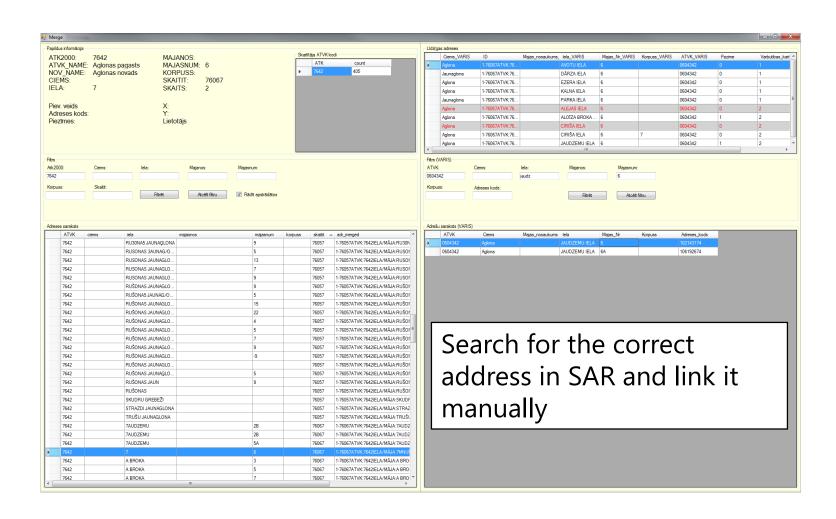
	Persons	% of population
Linked automatically	2 080 803	88
Linked manually (incl., linked automatically without coordinates, more than one address linked automatically and errors corrected in automatic linking)	296 070	12
Address cannot be geocoded or existence is doubtful at the critical moment of the PHC	510	0

Central Statistical
Bureau of Latvia