

# Determining Permanent Residency Status Using Registers in Estonia

Kristi Lehto<sup>1</sup>, Ethel Maasing<sup>2</sup>, Ene-Margit Tiit<sup>3</sup>

<sup>1</sup> *Statistics Estonia, Tallinn, Estonia; kristi.lehto@stat.ee*

<sup>2</sup> *Statistics Estonia, Tallinn, Estonia; ethel.maasing@stat.ee*

<sup>3</sup> *Statistics Estonia and University of Tartu, Tartu, Estonia; ene.tiit@ut.ee*

## **Abstract**

The next population and housing census in Estonia at the end of 2020 is intended to be register-based. The Estonian Population Register is over-covered for the census. It can be assumed that the people who actually live in Estonia are represented in other administrative registers because they are using services and receive payments.

The paper gives an overview of the background of using registers to determining permanent residency status and development of residency index in Estonia. Theory of residency index is described in section two and results are given in last section. This method is also planned to be used in regular population statistics.

**Keywords:** Population, census, registers, residency index

## **1. Registers and population size in Estonia**

A number of residents or population size is important for all countries, but also cities, towns, and municipalities. For a long time the only way to get information about the number of residents was a census.

From the time when different registers were created and implemented, the situation has changed, as the number of residents can be counted also from registers. Since it seems that in countries having population register or some other good (administrative) registers, the population size can be calculated at any time without interviewing the people.

But in reality, the situation is not so simple. A number of sources of information sometimes complicates the situation because the results might be inconsistent. For instance, in Estonia

after population and household census 2011 (PHC2011) we had three different numbers of population size:

- Size of census population – 1 294 455.
- Population size calculated using registered population events and population size of census PHC2000 – 1 320 000.
- Number of Estonian residents in Estonian Population Register (PR) – 1 365 000.

### *1.1. Under-coverage of PHC2011 and estimating the true population size*

After PHC2011, it became evident that census population was somewhat under-covered. This situation is very common nowadays when the people are very mobile and the migration between countries belonging to EU and/or Schengen group is free. Also, it seemed that probably the population size fixed in PR was somewhat over-covered. In 2012, immediately after PHC2011, the true size of Estonian population was estimated (Tiit, Meres, Vähi, 2012; Tiit, 2012; Tiit, 2014).

For this aim, the set of people belonging to PR, but not enumerated in PCH2011 (60 000 persons, about 4.6% of the population) was investigated using the existing system of administrative registers including 12 registers. The activities of these 60 000 problematic persons in all registers during the year 2011 were checked. So for each person were created 12 binary variables demonstrating their activity in every register. The residency was estimated statistically, using these binary variables as explanatory variables for logistical and linear regression. For training groups served census data, where enumerated people formed the group of residents and emigrated by word of relatives were taken as non-residents.

About 30 000 persons (2.3% of population) were added to census population to get the official population for demographic calculations. Each added person was identified by his/her ID-code (more exactly: recoded ID-code that does not allow to identify person, but allows to put together all his/her data from different registers).

There were two main reasons why the census population and population of Estonian residents in PR differed. Population of Estonian residents in PR included non-registered emigrants who had left Estonia during more than 10 years and hence was over-covered. The same situation is common in many so-called transition countries. Census population was under-covered as people nowadays estimate very highly their privacy and hence are not very keen on participating censuses. This problem is common in most (developed) countries.

### *1.2.Preparation for PHC2020. Estimation of census population*

As Estonia has a quite well-functioning registers system, it has been decided that the following population and household census in 2020/2021 will be organised without personal enumeration and interviewing, but based on using registers, as this has been made already in Nordic countries, Austria, Slovenia, and Netherlands (Tiit, 2015). That means it is necessary to know beforehand the census population – the identified set of residents. For these people, all census variables will be collected and/or calculated by data gained from existing registers.

It is reasonable that the task of estimating (future) census population coincides with the current calculation of annual population: every year the population of the previous year is corrected adding immigrants and children born last year and subtracting emigrants and people dead last year. While the data of natural increase (births and deaths) is exact nowadays, then migration data might be quite inaccurate due to defective registration that has lasted for a long time. Hence, it will be complicated to include into a list of residents people who have left without registering and return some years later.

One possibility for solving the problem is to create the model for residency testing using all existing registers to build explanatory variables. The activity in registers depends on the sex and age of person. This fact was taken into consideration when estimating the under-coverage of PHC2011 (Tiit, Meres, Vähi, 2012; Tiit, 2012; Tiit, 2014) and preparing the residency models for register based census (Maasing, 2015). To use different age-groups and different models in calculating indexes every year were too troublesome. It is more reasonable to take into account the ratio of definite residents and definite non-residents in each register.

## 2. Residency testing using residency index

### 2.1. Principal concepts for formulating the task of residency testing

**Time.** The whole process of checking residency is connected with one fixed year. This fact follows from the common residency rule used in census statistics: a person attains (and also loses) residency of a country during a year. Hence, the residency status of a person in year  $k + 1$  is defined by his activities in year  $k$ .

**Persons.** Let us have the maximal population  $M$ , that is set for person  $j$ ,  $j = 1, 2, \dots, J$  about whom we have to make the decision if they are residents or non-residents. The content of maximal population changes every year: the people will be added if they immigrate (officially) or are born. The only feasible reason for dropping off from the population  $M$  is death.

**Registers.** Let us have a set of registers/subregisters  $i$ ,  $i = 1, 2, \dots, I$ . We assume that they are independent in the sense that the data from one register are not, in general, copied into another register of the set. To each person  $j$ , register  $i$  and year  $k$  a binary variable  $B(i, j, k)$  accords in the following way:

- $B(i, j, k) = 1$ , if the person  $j$  has been at least once active in register  $i$  during the year  $k$ ;
- else  $B(i, j, k) = 0$ .

### 2.2. Generalized sum of signs of life

Let us form for every subject  $j$  of the population  $M$  a linear combination of all binary variables reflecting the activity in registers in year  $k$ ,

$$X_j(k) = \sum_{i=1}^I a_i B(i, j, k), \quad (1)$$

where  $a_i$  are fixed parameters. This is called *generalized sum of signs of life*.

Name of signs of life was introduced by Zhang and Dunne (Zhang, Dunne, 2015).

The value  $X_j(k)$  may have different content depending on the concrete task.

- When  $k$  is fixed and all parameters  $a_i$  equal to 1, then  $X_j(k)$  is the so-called *simple sum of signs of life*.
- When  $k$  is fixed and parameters  $a_i$  are weight of register, then the value  $X_j(k)$  is the *weighted sum of signs of life*. See section 3.2.

### 2.3. Residency index

To avoid the instability of residency that might be caused by independently created yearly models and warranting stability of estimated resident population the idea of **residence index** has been launched. The main essence of the idea is using in maximal amount the results of preceding years in predicting the residency status of a person in a current year.

If in every year the residency status for each person from the set  $M$  will be defined independently of his/her status last year, then the definition process is substantial without memory. Such situation is not consistent with the content and meaning of the process in real life, as changing the residency status is for a people comparatively infrequent event. Our aim is to create a mechanism for defining residency for people  $j$  for year  $k$  that is more stable in consecutive years.

Let us assume that for all persons from population  $M$  their residency status for a year  $k$  has been fixed. Define for them the **residency index**  $R_j(k)$  in the following way:

- $R_j(k) = 1$ , if the person  $j$  is resident in the year  $k$ ;
- $R_j(k) = 0$ , if the person  $j$  is not a resident in the year  $k$ ;
- $0 < R_j(k) < 1$ , if the person  $j$  residency status is not clear.

By definition always hold the inequalities:

$$0 \leq R_j(k) \leq 1. \quad (2)$$

Hence,  $R_j(k)$  can be interpreted as (subjective) probability that subject  $j$  in year  $k$  is resident. To ensure the condition (2) the value of indicator  $R_j(k)$  always must be cut down to value 1 or cut up to value 0.

In practical decision process there exists also a threshold  $c$  ( $0 < c < 1$ ) so that if  $R_j(k) \geq c$ , then person  $j$  has been considered as resident in year  $k$ .

For calculation/assigning the value  $c$  there are some traditional rules in the case when  $R_j(k)$  has been defined using statistical models. In other cases the value of threshold  $c$  must be derived empirically.

#### *2.4. Recalculation of residency index*

The key question in defining the residency index is – how to calculate the residency index of all members of population  $M$  for consecutive years. Formally, let the break-down point be the beginning of year, the 1<sup>st</sup> January.

We assume at the beginning of year  $k + 1$  that most people from population  $M$  already have the index  $R_j(k)$  that should be recalculated. The only people who do not have the index are newcomers.

All people  $j$  who were added (births and immigration) to population  $M$  during the year  $k$  will have

$$R_j(k + 1) = 1. \quad (3)$$

In the case of immigrants it is not important if they enter the first time or have been residents also earlier.

And people  $j$  who are in population  $M$  and have registered their emigration during the year  $k$  will have

$$R_j(k + 1) = 0. \quad (4)$$

For other persons from  $M$  the most logical and simple way is to use the linear combination of two indicators from previous year – residency index  $R_j(k)$  and generalized sign of life  $X_j(k)$ :

$$R_j(k + 1) = d * R_j(k) + g * X_j(k). \quad (5)$$

Both parameters  $d$  and  $g$  must satisfy the conditions  $0 \leq d, g \leq 1$ .

It is obvious that the bigger is the value  $d$  the more stable is the process and the more likely the persons save their residency status from year to year. From combination of values  $d$  and  $c$  depends how long it takes that a resident will acquire the status of non-resident and vice versa. For instance, if the condition

$$d^q < c \leq d^{q-1} \quad (6)$$

holds, then it is possible that a resident loses the status of resident if she/he has sign of life permanently zero during  $q$  years.

On the contrary, if a person has a lot of signs of life she/he can get the residency status within one year.

### *2.5. Estimation of parameters*

The three parameters:  $c$ ,  $d$  and  $g$  defining the decisions, cannot be estimated statistically, as there are no additional information. Hence at least initial values should be estimated using some logical considerations.

Values  $c$  and  $d$  define the exclusion time, that is, the time how long can the person be resident without any signs of life, see formula (6). Inclusion time depends on parameters  $g$  and  $c$ , but also from distribution of  $X_j(k)$  that is influenced by weights  $a_i$  used.

In Estonia we chose that suitable values are:

- $c = 0.7$ ,
- $d = 0.8$  and
- $g = 0.2$ .

### 3. Using the residency index for estimation Estonian population

#### 3.1. Defining the initial population

The first step in defining the set of residents is fixing the initial maximal population  $M$ . This population should contain all people who principally might belong to the set of residents. In Estonia this was the population of (alive) people fixed in PR being either residents or not, but having Estonian ID-code. Also the population  $M$  included people who were enumerated in PHC2011, but were not Estonian residents in PR (number of such persons was vain). In future the size of population  $M$  may somewhat increase when also people fixed in other registers, but not in PR will be included to  $M$ .

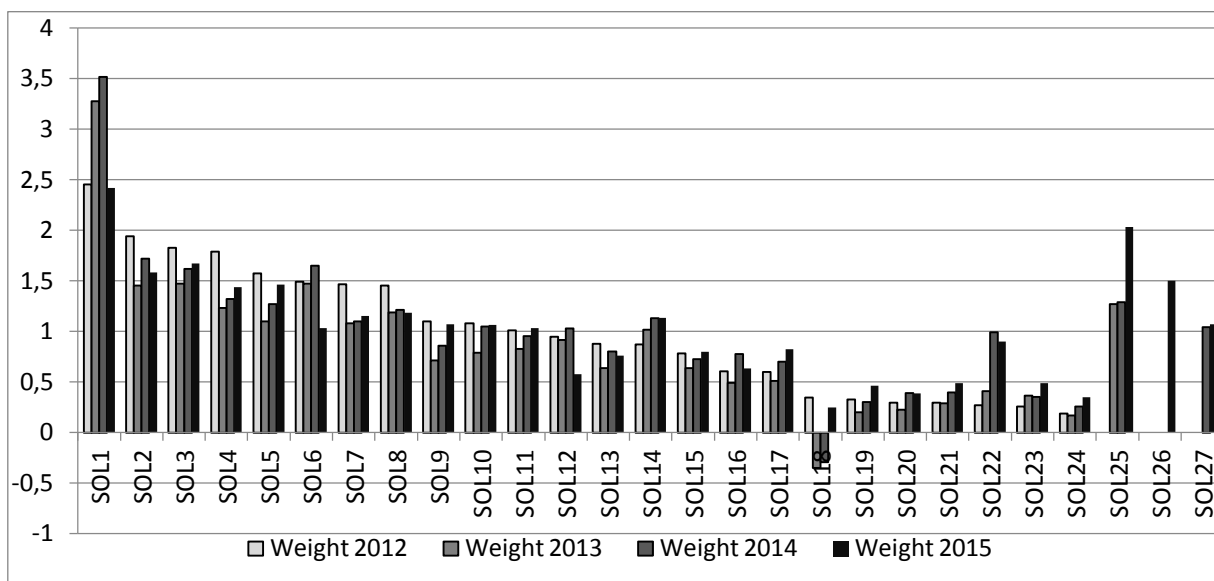
Every year the initial population changes: there will be a set of newborns and new (registered) immigrants who are added to initial population. All people dead will be deleted, but not emigrants – they will remain members of population  $M$ .

#### 3.2. Weighting signs of life

There are different ways to define the coefficients  $a_i$  in the expression (1).

- To take into account simply signs of life that is simply to summarize all register-based binary variables, that is to take all parameters  $a_i$  equal to 1. In Estonia this way did not show very good results.
- To calculate parameters  $a_i$  from some model. This idea needs special models for all sex-age-groups and might have the problem of inclusion non-typical residents. In Estonia we did not do this.
- To use weighted binary variables where weights are proportional to their ability to differentiate residents and non-residents. In Estonia we tried these weights and got better results than with simple sum of signs of life.
- Instead of ratios having quite wide amplitude of variability the logarithms of the ratios can be used as weights. These are the weights that work the best in Estonia (Fig.1).





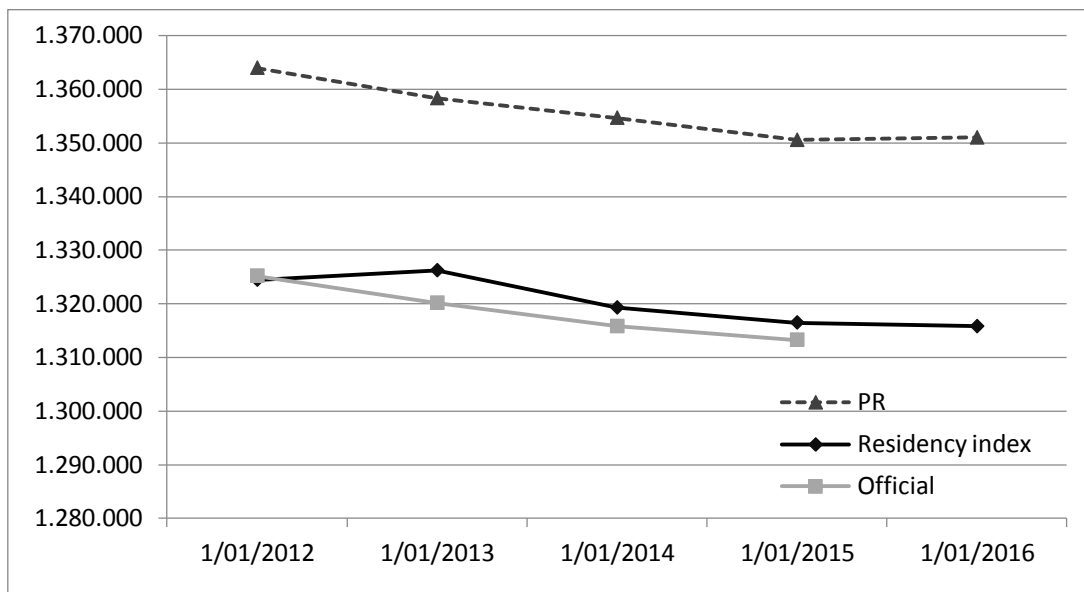
**Fig. 1** Logarithm of the ratio of definite residents and definite non-residents in all registers/subregisters

In total we can use 27 different signs of life from registers/subregisters in 2015, for example working in Estonia, changing or getting driver license, getting pension from Estonian government, different health care benefits and so on. Every year the new weights for registers/subregisters are calculated. Recalculation of weights gives the sustainability to residency index and takes into account of possible changes in the registers. It gives an opportunity to add or remove the registers/subregisters over years.

### 3.3. Estonian population by residency index

The result of residency index is shown in Fig.2. Starting point for calculating the residency index is 01.01.2012. For the first index values PHC2011 data were used. In first year, index result isn't very reliable, because of signs of life are known only for one year. Since 01.01.2014 results are very similar to official population number, the difference is less than 0.5%. The Estonian PR is over-covered, because of people who have left Estonia and do not have their leaving registered in the PR.

Method of residency index is also planned to be used in regular population statistics.



**Fig. 2** Estonian population by three different sources

#### 4. References

Maasing, E. (2015) Permanent Residency Status Determination in Register-Based Census, Master thesis, University of Tartu

Tiit, E.-M. (2012) Assessment of under-coverage in the 2011 Population and Housing Census, Quarterly Bulletin of Statistics Estonia, 4, pp. 110-119.

Tiit, E.-M. (2014) 2011 Population and Housing Census. Methodology

Tiit, E.-M. (2015) The register-based population and housing census: methodology and developments thereof, Quarterly Bulletin of Statistics Estonia, 3, pp. 48-71.

Tiit, E.-M., Meres, K., Vähi, M. (2012) Assessment of the target population of the census, Quarterly Bulletin of Statistics Estonia, 3, pp. 79-108.

Zhang, L.-C., Dunne, J. (2015) Census like population size estimation based on administrative data, Fourth Baltic-Nordic Conference on Survey Statistics.