STATISTICS IN TRANSITION new series, June 2017
Vol. 18, No. 2, pp. 211–226, DOI: 10.21307/stattrans-2016-067

RESIDENCY TESTING. ESTIMATING THE TRUE POPULATION SIZE OF ESTONIA

Ene-Margit Tiit[1]

ABSTRACT

The number of residents or population size is important for all countries. Nowadays in many countries a series of registers have been created, which can be used for assessing the population size. The residency index is a tool created for estimating the under- and over-coverage of population census and calculation of proper population size. For this aim the concept of a sign of life – a binary variable depending on register i, person j and year k has been introduced showing if the person was active in the register in a given year. The weighted sum of signs of life indicates the probability that the person belongs to the set of residents in a given year. To improve the stability of the index a linear combination of the previous value of the index and the sum of signs of life is used. Necessary parameters were estimated using empirical data.

Key words: population size, under-coverage of census, sign of life

Residency and population size. The case of Estonia

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

From the time when different registers were created and implemented, the situation has changed, as the number of residents can also be counted from registers. Therefore, it seems that in the countries that have a population register or some other good (administrative) registers the population size can be calculated at any time without interviewing the people [1].

In reality, however, the situation is not so simple. Multiple sources of information sometimes complicate the situation because the results may be inconsistent. For instance, in Estonia after the population and household census of 2011 (PHC2011), we had three different numbers of population size:

- The size of census population – 1 294 455.

[1] University of Tartu, Faculty of Science and Technology. E-mail: ene.tiit@ut.ee.

- Population size calculated using registered population events and the population size of census PHC2000 – 1 320 000.
- The number of Estonian residents in the Estonian Population Register (EPR) – 1 365 000.

In some age groups, the difference between various estimates was almost 10%.

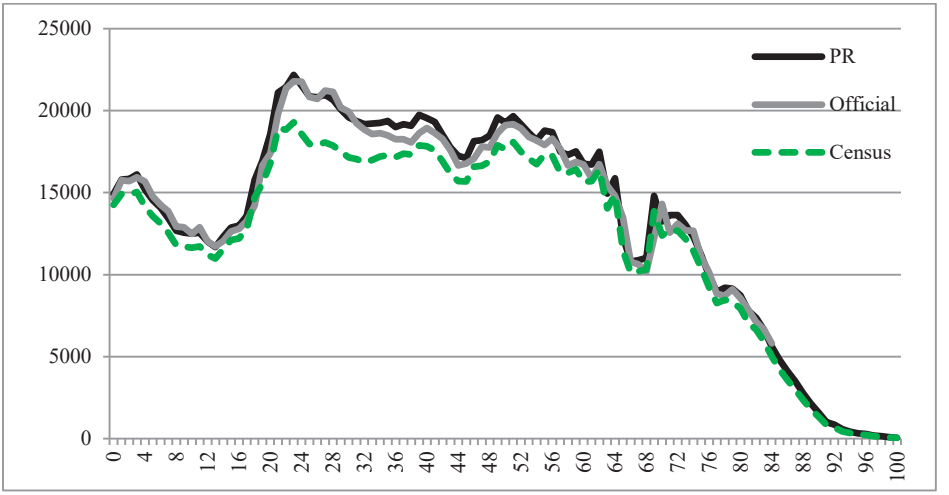

Figure 1. Age-sex distribution of the Estonian population by three different sources

Under-coverage of PHC2011 and estimating the true population size

After PHC2011, it became evident that census population was under-covered. This situation is very common nowadays when the people are very mobile and migration between the countries belonging to the EU and/or the Schengen group is free. It also seemed that probably the population size fixed in EPR was over-covered. In 2012, immediately after PHC2011, the true size of the Estonian population was estimated, see [2—4].

For this aim, the set of people belonging to EPR, but not enumerated in PCH2011 (60 000 persons, about 4.6% of the population) was investigated using the existing system of administrative registers, which includes 12 registers. The activities of these 60 000 problematic persons during the year 2011 were checked in all registers. Thus, 12 binary variables demonstrating their activity in every register were created for each person. Residency was estimated statistically, using these binary variables as explanatory variables for logistical and linear regression. For completing training groups needed in statistical procedures the census data were used. For different age-sex groups different models were created, as the activity in registers depends on age, see Figure 2, where for each person from the training group the sum of all binary variables is presented.

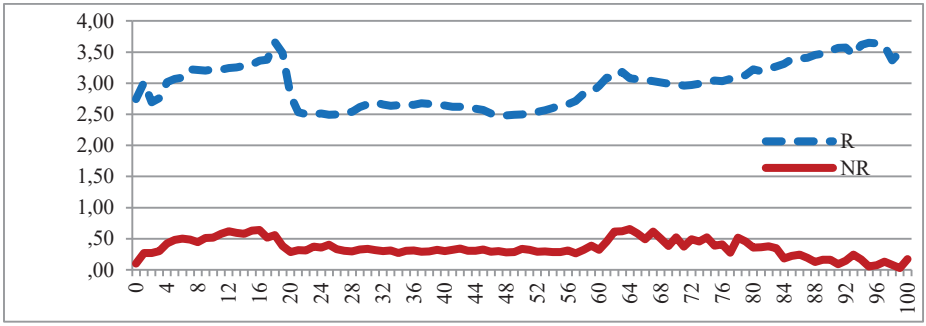

Figure 2. Total activity of residents and non-residents depending on age

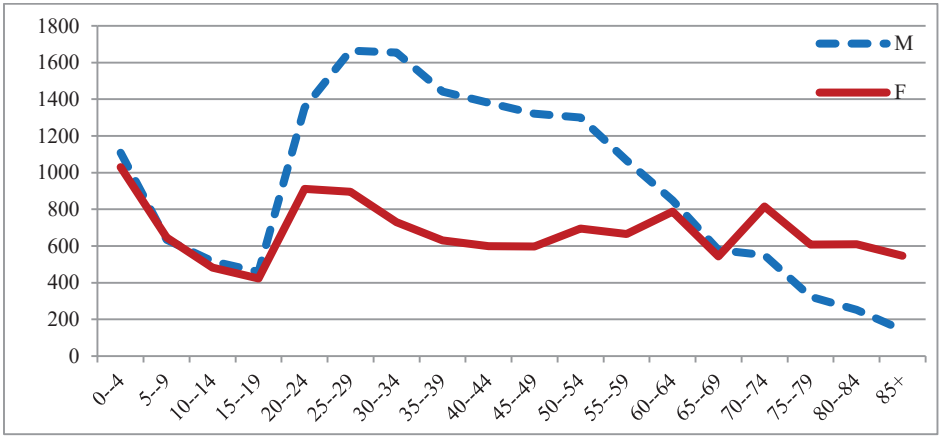

Figure 3. The sex-age distribution of estimated residents (under-coverage of census) added to the census population

About 30 000 persons (2.3% of population) were added to the census population to get the "official" population which Statistics Estonia was using for demographic calculations. In all sex-age groups, the inclusion and exclusion errors of the model were less than 5%. Each added person was identified by his/her recoded ID-code. Such codes allow combining the person's all data from different registers without identifying the person.

There were two main reasons why the census population and the population of Estonian residents in EPR differed. The population of Estonian residents in EPR included non-registered emigrants who had left Estonia during more than 10 years, and hence it was over-covered. The same situation is common in many other transition countries. The census population was under-covered as people are very mobile nowadays, and they also appreciate their privacy very highly, and therefore, are not very keen on participating in censuses. This problem is common in most European countries [5].

Preparation for PHC2020. Estimation of census population

As Estonia has a quite well-functioning system of registers, 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 registers, as this has already been done in the Nordic countries, Austria, Slovenia and the Netherlands [6]. That means it is necessary to know the census population – the identified set of residents – beforehand. All the census variables about these people will be collected and/or calculated by the data gained from the existing registers.

It is reasonable that the task of estimating the (future) census population relies on current calculation of annual population: every year the population of the previous year is corrected via adding the immigrants and the children born that year and subtracting the emigrants and the people who died that 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 decades. Due to errors made in the past, it is complicated to include into the list of immigrants people who have left without registering and returned after some years.

One possibility is to create the model (similar to the model of estimating under-coverage) for residency testing using all the existing registers as explanatory variables. In this case the following problems arise:

- Who are the people to be checked?
- How to get reliable training groups?
- Are all registers equally important, reliable and also independent?

One attempt to solve this problem was made in 2015, three years after the first correction of the "true" population [7, 8]. The solution was the following:

Training groups were formed using information of PHC2011, the current population events and EPR. The population to be checked consisted of the total EPR population (residents and non-residents). All the functioning registers (21 registers and sub-registers) were used to create explanatory variables. Decisions were made using logistic regression. The results were also checked with the help of linear regression and discriminant analysis.

The results of the attempt were positive in the sense that the ideology worked. It was possible to create explanatory variables by registers and check the residency of people using the statistical model. But some problems also occurred.

1. The estimated number of residents (population size) was about 5% lower than the currently calculated official "true" population size;
2. There was a number of indeterminable people who had a different pattern of registers' activities compared with typical residents and typical non-residents; see Figure 4, where on the right, there is the cluster of residents, on the left, the

column of non-residents, and between them, the series of indeterminable subjects.

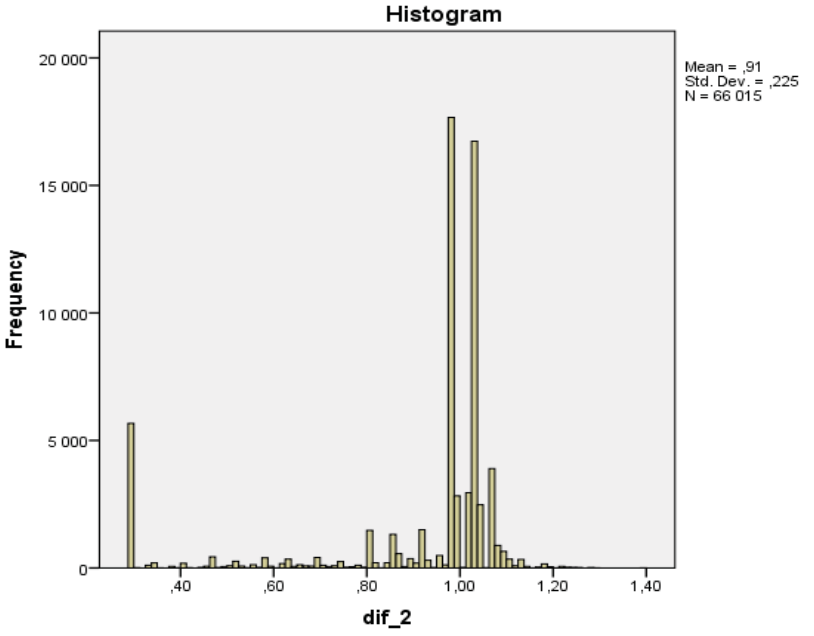

Figure 4. The histogram of distribution of values ascribed to people by the discrimination model

3. There exist a number of persons who are residents but do not show any activity in registers. In the population of PHC2011, there were about 3% of people who had not been active in 2011 in any of the 12 registers used for testing under-coverage. Such people were mainly working-aged men who did not study, did not visit doctors, did not get any social support and, probably, did some non-registered (the so-called black or grey) work.
4. If the strategy is to make a decision separately for each year using new models every time, then it is difficult to gain stability of population (on the personal level). Conversely, when the same model has been used for several years, it will lose optimality.

To avoid the problems listed, a new approach – using the residency index – was invented [9, 10].

Residency index

Principal concepts for formulating the task of residency testing. Fuzzy sets

Time. The whole process of checking residency is connected with one fixed year. This fact follows from the tradition of assessing the population number at the

beginning of the year. The common residency rule used in census statistics also has the lag-time of one year – a person attains (and also loses) the residency of a country during a year. Hence, the residency status of a person in year k is defined by his activities in year k—1.

Persons. Let us have maximum population M, that is a set of persons j, j=1, 2, …, J about whom we have to make the decision if they are residents or non-residents. The content of maximum population changes every year – people will be added to M if they immigrate or are born. The only feasible reason for dropping off from population M is death.

Registers. Let us have a set of registers/sub-registers i, i = 1, 2, …, I. We assume that they are independent. For each person j, register i and year k, a binary variable $B(i,j,k)$ is defined in the following way:

$$B(i,j,k) = \begin{cases} 1, \text{ if person } j \text{ has at least once been active in register } i \text{ in year } k \\ 0 \text{ else.} \end{cases} \quad (1)$$

We say that $B(i,j,k)$ is the ***sign of life*** (SL) of person j in year k. This term has been introduced by Li-Chun Zhang and Dunne [10].

Summarised SLs

Let us form, for every subject j of population M, a linear combination of all binary variables SL reflecting his/her activity in registers in year k,

$$X_j(k) = \sum_{i=1}^{I} a_i \, B(i,j,k), \quad (2)$$

where a_i are fixed coefficients. The value $X_j(k)$, the generalised sum of SLs, may have a different content depending on the concrete task and values of weights.

1. When k is fixed and all the parameters are equal to 1, then $X_j(k)$ is the ***simple sum of*** SLs.
2. When k is fixed and parameters a_i are calculated as coefficients of the discriminating model (e.g. linear or logarithmic regression), then the value X_j is the prognosis of the residency status of subject j, see [7].

Residency index

To assure the stability of the estimated resident population, the idea of the ***residency index*** has been launched. The main essence of the idea is to predict the residency status for all potential residents every year, using the whole information about them collected during the preceding years.

Assume that for all persons from population M their residency status for year k has been fixed and define the **residency index $R_j(k)$** for them in the following way:

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

By definition, the inequalities (3) always hold:

$$0 \leq R_j(k) \leq 1. \tag{3}$$

The residency index can also be treated using the concept of fuzzy sets introduced by Zadeh and Klaua [11, 12]. In this framework, R is for a given year k a membership function from the population M, $R: M \rightarrow [0.1]$ and for each person j the $R_j(k)$ is the grade of membership in year k. $R_j(k)$ can also be interpreted as (subjective) probability that subject j is a resident in year k. To ensure the condition (3), the value of indicator $R_j(k)$ must always be truncated.

In the practical decision process, not only the people having the residency index equal to 1, but also some others belong to the set of residents. That means there exists 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. (4)

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 general, the value of threshold c must be derived considering rational calculations, and their consonance with empirical data should be tested statistically. In the following, we say that residents having $R_j(k) = 1$ are ***confident residents*** (CR) and non-residents having $R_j(k) = 0$ are ***confident non-residents*** (CNR).

Recalculation of the residency index

The key question in defining the residency index is – how to calculate the residency index for all members of population M for consecutive years? We assume at the beginning of year k+1 that most people from population M have the index $R_j(k)$ from the previous year that should be recalculated. The only people who do not have the index are newcomers. All people j who were added to population M during year k will have

$$R_j(k+1) = 1. \tag{5}$$

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

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

$$R_j\,(k+1) = d\,R_j(k) + gX_j(k). \tag{6}$$

Both the ***stability parameter*** d and the ***SL parameter*** g must satisfy the conditions $0 \le d,g \le 1$. As term $X_j(k)$ is not restricted by 1, there is no need to use the convexity condition $d + g = 1$. To ensure the condition (3) the value $R,(k+1)$ is truncated:

$$\text{If } R_j(k+1) > 1, \text{ then } R_j(k+1) := 1. \tag{6a}$$

Estimation of parameters

The three parameters – c, d and g defining the residency of persons are connected with the following decisions:

- How long a CR can stay in the status of a resident without any SL? This is the ***exclusion time*** q_1.
- How long does it take for a CNR to obtain residency status on the basis of SLs? This is the ***inclusion time*** q_2.

Parameters *c* and *d* and exclusion time

As regards the first question, we can see that the bigger the value *d,* the more ***stable*** the process, and the more likely the persons are to retain their residency status for a longer time. Exclusion probability and exclusion time also depend on the value of *c*: the higher the value *c*, the more probable it is that a person *j* will be excluded from the set of residents. We can also say that the higher *c*, the more ***conservative*** the decision.

Hence, it depends on the combination of values of *d* and *c* how long a person will retain the status of a resident not having any SL. If the condition (7) holds, then the CR having no SL retains the status of a resident for *q*—1 years, and loses it after that.

$$d^q < c \le d^{q-1}. \tag{7}$$

From (7) it follows that the change of residency happens at the moment

$$q = \frac{\ln c}{\ln d}. \tag{8}$$

As the recalculation of index happens at the beginning of the year, the exclusion time is the smallest integer q_1 satisfying the condition

$$q_1 \ge q.$$

Parameter g and inclusion time

To analyse the second question, we have to pay attention to CNRs who are obtaining residency status using SLs. Here, we assume that a_i =1. Then, the condition that the person obtains residency exactly in q years having every year f SLs is the following:

$$fg(1 + d + d^2 + \cdots + d^q) \geq c > fg(1 + d + d^2 + \cdots + d^{q-1}).$$

As the brackets contain the sum of geometric progression, we get the inequalities for parameter g, see (9):

$$\frac{c(1-d)}{f(1-d^{q+1})} \leq g < \frac{c(1-d)}{f(1-d^{q})}. \quad (9)$$

The necessary conditions for positive probability of getting residency status by SLs follow from the sum of geometric progression:

$$g \geq \frac{c(1-d)}{f}. \quad (10)$$

From inequality (9), we get the expression for inclusion time:

$$q = \ln\left\{\frac{fg-c+cd}{fg}\right\} / \ln d - 1 \quad (11)$$

q_2 is the smallest integer fulfilling the condition $q_2 \geq q$, defined by (11).

The special case when $fg \geq c$, then q_2=1 means the person gets the status of R in the first year.

Table 1. Exclusion and inclusion time in the case of a selected set of parameters

Case No	c	d	g	f	q_1	q_2
1	0.7	0.75	0.2	1	2	7
2	0.7	0.75	0.2	2	2	1
3	0.7	0.75	0.25	1	2	4
4	0.7	0.75	0.25	2	2	1
5	0.7	0.8	0.2	1	2	5
6	0.7	0.8	0.2	2	2	1
7	0.7	0.8	0.25	1	2	3
8	0.7	0.8	0.25	2	2	1
9	0.75	0.75	0.2	1	2	9
10	0.75	0.75	0.2	2	2	2
11	0.75	0.75	0.25	1	2	4
12	0.75	0.75	0.25	2	2	1
13	0.75	0.8	0.2	1	2	6
14	0.75	0.8	0.2	2	2	2
15	0.75	0.8	0.25	1	2	4
16	0.75	0.8	0.25	2	2	1

From the table we can see that the exclusion time is quite stable (and does not depend on f and g), while the number of SLs has a big influence on the inclusion time. For the following example we will choose the parameters c=0.7, d = 0.8 and g= 0.2. That means the exclusion time is 2 years (the CR will be excluded from the set of Rs after two years without SLs), and inclusion time in the case of f= 1 and 2 is correspondingly 5 and 1. Hence, a CNR will gain the residency status in one year if s/he gets 2 SLs and in five years getting one SL each year. The last assertion is true in the case of the simple sum of SLs. When the weighted sum SLs is used, then this calculation is true in average. That means if the SLs have weights that differ from the mean weight, the inclusion time might be somewhat different: using a SL having low weight, the inclusion time might be longer, and in the case of SLs having high weight, the inclusion might be faster.

Example. Using the residency index for the estimation of Estonian population

Maximal population model parameters and initial residency index

The first step in defining the set of residents is fixing the initial maximum population M. This population should contain all people who, in principle, might belong to the set of residents. In Estonia this set is the population of (living) people fixed in EPR, being either residents or not but having an Estonian ID-code. Population M also includes people who were enumerated in PHC2011 but were not Estonian residents in EPR (the number of such persons was very small). In the future, the size of population M may somewhat increase when people also fixed in other registers but not in EPR will be included in M.

The model parameters will be defined in the following way: d = 0.8, c = 0.7 and g = 0.2. Then, the model is rather conservative, and it takes several years to obtain residency by SL. For instance, a CNR can get the status of a resident by having one SL only during ten years. To get the status of a resident in one year, the person must have 5 SLs.

We will define the initial residency index R0 in the following way: using the fact that the critical moment of PHC2011 in Estonia was 31.12.2011, it almost coincides with the beginning of the year 2012.

- R0 = 0 for persons who were not Estonian residents in EPR on 1.01.2015 and were not enumerated in PHC2011;
- R0 = 1 for persons who were Estonian residents by EPR on 1.01.2015 and either were enumerated in PHC2011 or were born in 2012–2014.
- R0 = 0.8 for persons who were Estonian residents by EPR on 1.01.2015 but were not enumerated in PHC2011 and were added to the Estonian population using residency criteria in 2012 (the so-called under-coverage).
- R0 = 0.7 for all persons who officially immigrated in 2012–2014.
- R0 = 0.5 for all other persons from population M.

Table 2. Distribution of index R0

Population group	CNR	unclear	immigrants	under-coverage of PHC2012	CR	Total
Index	0.0	0.5	0.7	0.8	1.0	
Frequency	78 387	69 111	20106	25 094	1 270 161	1 462 859

Using the threshold 0.7, we have the number of residents 1 315 361, of which 1 270 161 are CRs.

Weighting SL

There are different ways to define coefficients a_i in expression (1). The activity in registers depends on the sex and age of person, see Figure 5. This fact was taken into consideration when preparing the models [3. 7], but it would be too troublesome to use different SLs for different age groups in calculating indexes, and it would cause instability of the processes.

It is common for all age-sex groups that the average activity of CRs and CNRs in registers is quite different, see Figure 8. But there are still differences between registers. Non-residents are more active in registers connected with health services (which are considerably cheap in Estonia) and also in the pension register.

From here it follows that it is reasonable to weight the register-based SLs taking into account the popularity of registers among residents and non-residents. Figure 7 depicts the ratio of activity of residents and non-residents in all the registers.

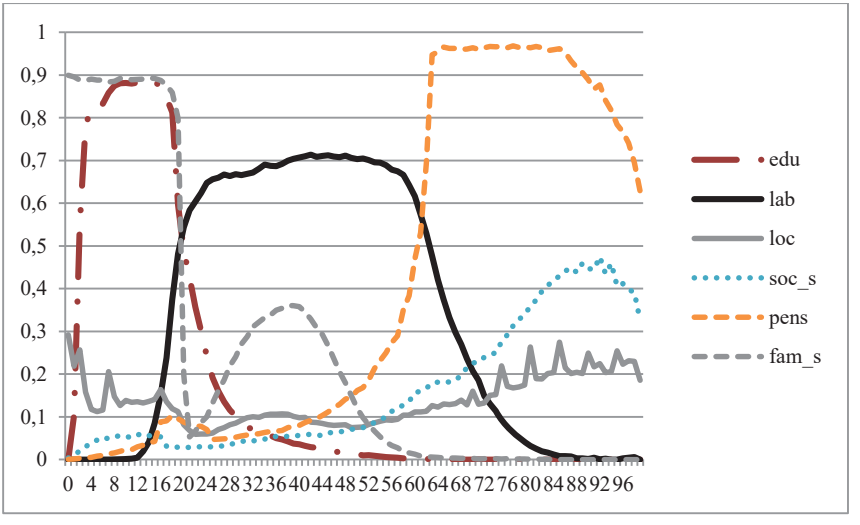

Figure 7. Average activity of all persons from *M* in registers depending on age.

Explanation of the names of variables: edu – learning in an Estonian school; lab – working in an organisation situated in Estonia; loc – getting any support from local administration; soc_s – getting social support or stipend from government; pens – getting pension; fam_s – family support, children benefit.

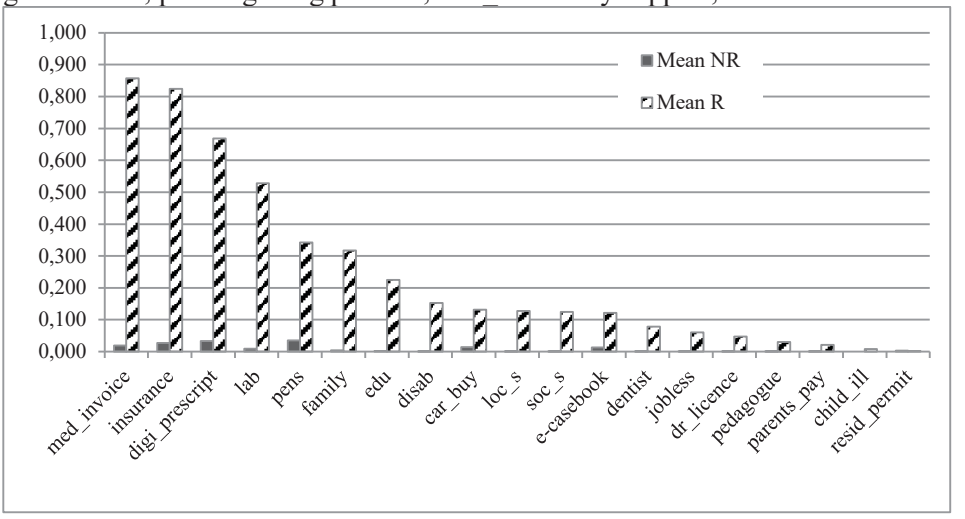

Figure 8. Average activity of CRs and CNRs in different registers.

Explanation of the names of variables: med_invoice – invoice for medical procedure; insurance – health insurance; digi_prescription – digital medical prescription; family – family support; disab – disability fixed by doctor; car_buy – buying or selling a car; e-casebook – fixed event in e-casebook; dentist – visiting dentist; joblessness – active in register of unemployed; de_licence – having doctot's licence; pedagogue – working as pedagogue; parents_pay – getting parents' support; child_ill – document of child's illness; resid_permit – residency permission;

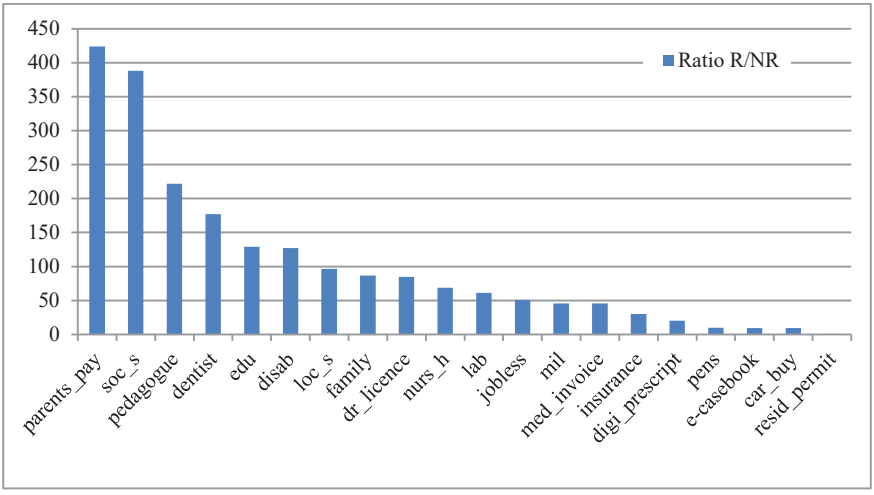

Figure 9. Ratio of average activities of CRs and CNRs in all registers

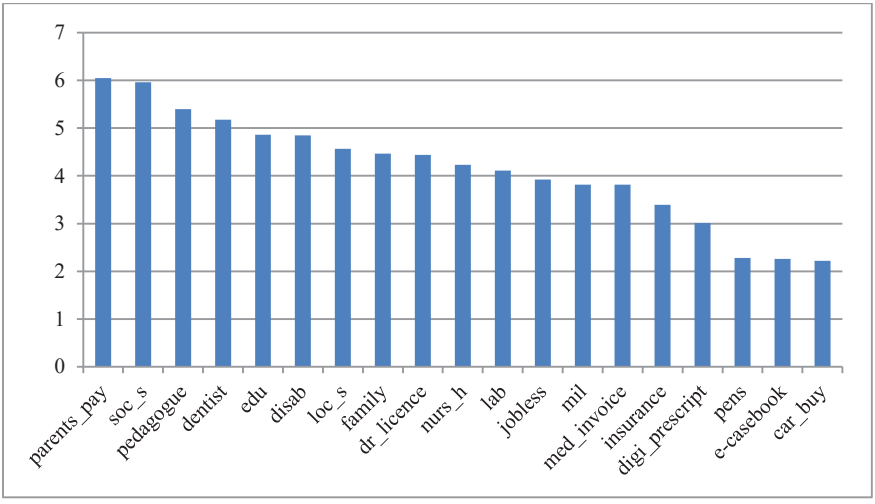

Figure 10. Logarithm of the ratios of activities of CRs and CNRs in all registers.

In defining GSL (see formula (2)), the following options will be used:

1. To take into account simply the SLs, that is to consider all parameters a_i equal to each other.
2. To use weighted SLs where weights are proportional to their ability to differentiate CRs and CNRs measured by ratios of average SLs in CRs and CNRs, see Figure 9.
3. Instead of the ratios of average SLs, the logarithms of the ratios are used as weights, see Figure 10.

In the following we will use and compare these options of calculating the weights. To ensure their comparability, all weights are normed by the mean of simple sum of SLs that equals 4.068, see Figure 11.

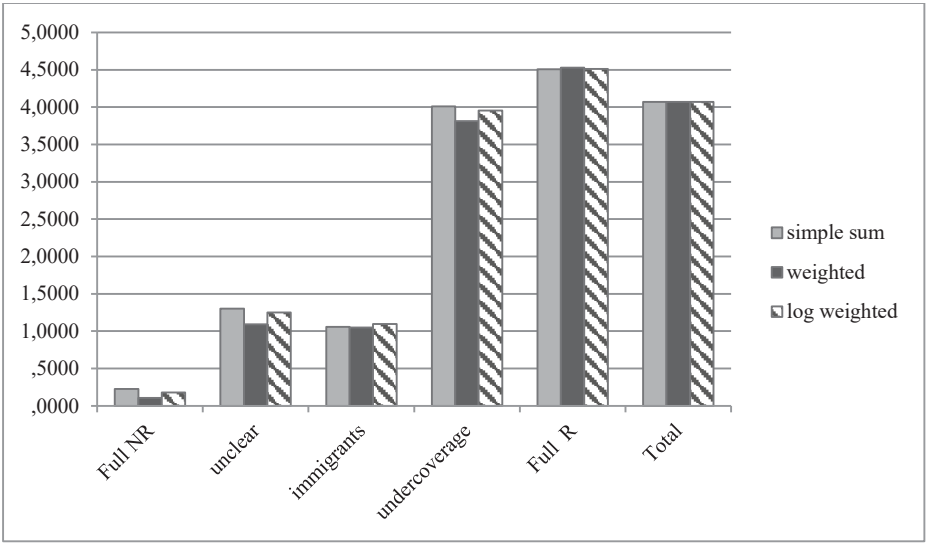

Figure 11. Weighted sum of SLs in different subpopulations of *M* (see Table 1).

It follows that the differences are not big, but the weighted sum is the most sensible.

Calculation of the residence index R1 for the next year

For calculation of the value of the index for the next year, formula (6) was used with parameters c=0.7, d= 0.8 and g = 0.2, and three different sets of weights. Table 2 gives the result of the decision in each case, the number of residents and non-residents, also the percentages.

Table 3. The number of residents and non-residents in the case of different weights for signs of life

	Number of residents	%	Number of non-residents	%	Increase in the number of residents
Simple sum	1 325 258	90.6	137 601	9.4	9 897
Weighted	1 318 385	90.1	144 474	9.9	3 024
Log weighted	1 318 585	90.1	144 274	9.9	3 224

In all the cases, the number of residents has increased by 0.2—0.7 %. In the case of weighted SL, the result is the closest to the supposed real situation.

Conclusion

The residency index is a tool for estimating the residency status of a single person from a population, therefore it can be used for estimating the coverage of a population census and also the population size of a country in an arbitrary year. The residency index uses the so-called signs of life that demonstrate the activity of a person in different registers. The most efficient is the calculation of the residency index in consecutive years, which gives a tool for monitoring the changes of population.

Compared with other types of statistical models, the advantage of this methodology is the possibility of using a large number of different registers, which may have both positive and negative impact on the residency status.

However, the use of indexes might be restricted by the specialities of registers. If the registers are not connected with detailed addresses of living places of persons, then residency indexes cannot be used for describing interior migration and population in small areas.

REFERENCES

Register-based statistics in the Nordic countries - Review of best practices with focus on population and social statistics (United Nations publication). http://unstats.un.org/unsd/censuskb20/KnowledgebaseArticle10220.aspx.

TIIT, E.-M., (2012). 2011. aasta rahva ja eluruumide loenduse alakaetuse hinnang. Eesti Statistika Kvartalikiri, 4/12, Quarterly Bulletin of Statistics Estonia, 4. 12, pp. 110–119.

TIIT, E.-M., MERES, K., VÄHI, M., (2012). Rahvaloenduse üldkogumi hindamine. Eesti Statistika Kvartalikiri, 3, pp. 79–108.

TIIT, E.-M., (2014), 2011. aasta rahva ja eluruumide loendus, Metoodika, 76+19 lk.

Main Results of the UNECE-UNSD Survey on the 2010 Round of Population and Housing Censuses (ECE/CES/GE.41/2009/25).

TIIT, E.-M., (2015). The register-based population and housing census: methodology and developments thereof. Quarterly Bulletin of Statistics Estonia. 3, 15, pp. 42–64.

MAASING, ETHEL, (2015). Eesti alaliste elanike määratlemine registripõhises loenduses**.** http://dspace.utlib.ee/dspace;/handle/10062/47557.

MAASING, ETHEL, (2015). First results in determining permanent residency status in register-based census, banocoss2015/Presentations?preview=#!/preview/149296295/170626623/ Maasing _Abstract.pdf.

TIIT, E.-M., (2015). Residence testing using registers – conceptual and methodological problems, https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=#!/preview/ 149296295/170626640/Tiit_Abstract.pdf.

LI-CHU, ZHANG, JOHN, DUNNE, Census like population size estimation based on administrative data https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=/149296295/172987273/CensuslikePopulationSize.pdf.

ZADEH, L. A., (1965). "Fuzzy sets", Information and Control, 8 (3), pp. 338–353.

KLAUA, D., (1965) Über einen Ansatz zur mehrwertigen Mengenlehre. Monatsb, Deutsch, Akad. Wiss, Berlin 7, pp. 859–876.