ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 21, Number 1, June 2017
Available online at http://acutm.math.ut.ee

Residency index – a tool for measuring the population size

Ethel Maasing, Ene-Margit Tiit, and Mare Vähi

Abstract. After the Estonian census 2011 the census team found that there was some under-coverage of the census data. To determine the amount of non-enumerated people the following procedure was used. The set of people belonging to Estonian population register as residents, but not enumerated in census 2011 were regarded as potential residents. All existing administrative registers were used to define the signs of life for these people: activity in a register during 2011 gave to a person a sign of life. The signs of life were used as binary variables to discriminant the residents and non-residents. The following task was to use the methodology for following years and to cover the whole population. Hence we decided to define for each person from the population a residency index between 0 and 1 that will be recalculated yearly using the signs of life.

From the very beginning of scientific thinking censuses have been the most important and valuable source of information on the number of residents (population size) of a state. Nowadays, when there exist different information sources, the reliability and exactness of census data does not satisfy. There are several reasons why the coverage of census has fallen – the mobility of people has increased, unwillingness to disclose personal data has arisen. Also, needs of researchers are higher today. Especially problematic situation occurs when it has been planned to carry out instead of traditional census a register-based census, as usually different registers have different list of residents and the real population size is unknown.

In Estonia the problem arose after the census of 2011, when we had three different population sizes: the number of residents listed in Population Register (1 365 000), the population size calculated in traditional way from the census 2000 data (1 330 000) and census 2011 data (1 295 000) [1]. It became

Received November 4, 2016.
2010 *Mathematics Subject Classification.* 62H05; 62H12; 91D20.
Key words and phrases. Distribution; estimation; residency index.
http://dx.doi.org/10.12697/ACUTM.2017.21.09

clear that census was under-covered [4]. So the first task was to estimate the size of under-coverage of census 2011 [4, 7].

Estimation of under-coverage of census 2011

As a source of information for estimation the administrative registers were used. In Estonia there were more than 10 administrative registers covering different fields of activities: education, health care, social support etc. The number of registers is, in general, increasing. All Estonian inhabitants – including people having temporary living permissions – have the Estonian ID-codes and are listed in Population Register (either as residents or as non-residents) [2, 3, 5]. All other Estonian administrative registers use the ID-codes for identifying persons. This makes all Estonian registers consistent. In fact, for making statistics not the real ID-codes but their encrypted versions are used so that statisticians cannot see the personal data of any person.

Signs of life

For using the registers as data sources the binary variables $E_i(j)$ are defined for each person ($j = 1, 2, \ldots, N$), and each register ($i = 1, \ldots, m$) in the following way:

$$E_i(j) = \begin{cases} 1, & \text{if the person } j \text{ has been active in the register } i, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The variables $E_i(j)$ are called signs of life [6], abbreviated SOL.

The values of SOL were used as explanatory variables to discriminate the residents and non-residents using several multivariate technics – linear and logistic regression and discriminant analysis. As training groups were used the **confident residents** who belonged to Population Register (PR) as Estonian residents and were enumerated in 2011 and the **confident non-residents** who were in PR as residents of some foreign country and were not enumerated as people living currently in Estonia in 2011. To get more adequate description the population was divided into groups by age and sex and special models were created for different groups. As a result, the under-coverage of about 2.3 % was found. The people from this group were identified by their recoded ID-codes. The error of decisions was also estimated, it was not more than 5 % in each group, totally about 3000 persons. The improved population list and size were accepted in making official population statistics in Estonia since 2012.

Estimation of population size three years later (2015)

All problems connected with non-registered migration continued after census 2011. The next step was estimation the whole population size in 2015 (after 3 years) using values of SOL and similar methodology. With this aim

the **enlarged population** was defined. It included all persons belonging to Estonian Population Register with permanent living place either in Estonia or somewhere else or without living place at all. The size N of enlarged population was about 1.5 millions, that is almost 15% more than estimated factual population. The models similar to the models elaborated for estimation the under-coverage of census 2011 (linear and logistic regression and discriminant analysis, in all age-sex groups) were used. But all models gave somewhat underestimated population size. The difference between the population sizes calculated in traditional way and estimated by the models was 1.6–4% (depending on statistical method used in the model).

Residency index

As it follows, a new approach was needed for estimating the residency status of population members. It is clear that values of SOL form a good tool to estimate the residency status of potential residents. But it is necessary to use also the information about the earlier residency status of persons. The solution should be efficient, as building a set of models for different groups of population is quite troublesome if it must be done every year.

The aim of our research was to build a common model for all sex-age-groups which uses values of SOL and also information from the past residency status of persons. It is also desirable that the methodology allows to use some regulations predicted by population policy of the country.

For each person j from the enlarged population ($j = 1, \ldots, N$) we defined the **residency index** $R(j,k)$ showing the probability that the person j in year k is resident. The index $R(j,k)$ has values in the interval [0, 1]. The limiting values of $R(j,k)$ have special meaning. When $R(j,k) = 0$, then j is a **confident non-resident** in the year k and when $R(j,k) = 1$, then j is a **confident resident** in the year k.

Every year the residency index is calculated for all persons from the enlarged population. The residency index for year k is calculated using the value of residency index in year $k-1$ and values of SOL gathered in year $k-1$:

$$R(j,k) = dR(j,k-1) + gX(j,k-1), \tag{2}$$

where d and g are parameters, and

$$X(j,k) = \sum_{i=1}^{m} a_i E_i(j,k), \quad a_i > 0, \tag{3}$$

is the sum of weighted values of SOL gathered in year k, see (1).

To ensure the condition

$$0 \le R(j,k) \le 1, \tag{4}$$

$R(j,k)$, calculated by formula (2), is truncated by values 0 and 1.

Additionally some conditions, natural in demographic calculations, are included: all persons born or immigrated in year $k-1$ get for the index $R(j,k)$ value 1, all persons dead or emigrated in year $k-1$ will have the index value 0 in year k.

Defining the model parameters

The **stability parameter** d and **SOL parameter** g in (2) have the values between 0 and 1,

$$d+g=1. \tag{5}$$

If $g=0$, then the formula (2) is the traditional formula of population statistics. If $d=0$, we get the model-based formula used in estimating the under-coverage and population size in 2015, where weights a_i are determined by the statistical procedure used.

To make decisions there must be fixed a **threshold** c $(0 \le c \le 1)$ so that

$$\text{if } \begin{cases} R(j,k) \ge c, & \text{then } j \text{ is resident in year } k, \\ R(j,k) < c, & \text{then } j \text{ is nonresident in year } k. \end{cases} \tag{6}$$

The values of parameters d and g and also the threshold c are defined in such way that some political conditions on saving and getting residency were warranted. These conditions are following.

Residency saving condition. The confident resident can save the residency status without getting any SOL during one year $(E_i(j)=0, \forall i)$, but not longer. That means, if confident resident has got no SOL during two years, he or she will became the status of non-resident.

Residency getting conditions.

I. The confident non-resident can get the residency status only during six years if he gets every year one SOL.

II. If a confident non-resident gets in one year at least five signs of life, he gets the residency status already in the next year. Five signs of life exceed the average number of values of SOL gathered by a resident in a year.

From these conditions and basic formula (2) the following inequalities connecting the parameters follow.

From residency saving condition the inequalities (7) follow

$$d^2 \le c \le d. \tag{7}$$

From the residency getting condition **I** arise inequalities

$$g\sum_{0}^{4} d^i < c \le g\sum_{0}^{5} d^i,$$

from which, using the properties of geometrical progression, follow the inequalities

$$1 - d^4 < c \leq 1 - d^5. \tag{8}$$

From the residency getting condition **II** we have the following condition:

$$5gd \geq c. \tag{9}$$

Calculation of admissible values of parameters

The task is to find the area on the plane of values of parameters d and c, where the conditions (7) – (9) are fulfilled. When this area exists and is found, then also values of parameter g can be easily calculated from the condition (5).

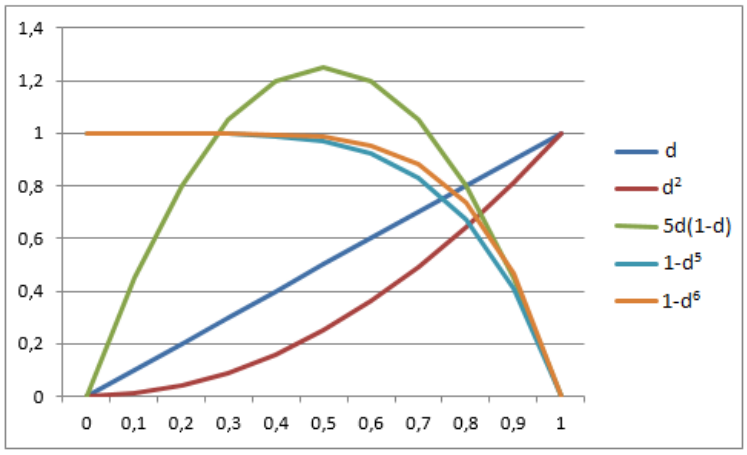

Figure 1. The areas satisfying given conditions (7) – (9).

The admissible region for c in the sense of conditions (7) is between the straight line $c = d$ and quadratic parabola $c = d^2$. The admissible region in the sense of conditions (8) is between the parabolas of 5^{th} and 6^{th} degree. The intersection of these areas is the small curvilinear rectangle with angles (0.7549; 0.7549), (0.8087; 0.6540), (0.8255; 0.6815) and (0.7781; 0.7781), see Figure 1.

The coordinates of angles are found via solution of systems of equations from the conditions (7) – (9).

The condition (9) illustrated by the big quadratic parabola with peak at point (0.5; 1.25) does not add restrictions. From the calculations it follows that possible values for d lie between 0.755 and 0.8255 and possible values for c lie between d and d^2, must be between 0.654 and 0.778.

In all following calculations we will use the values $d = 0.8$, $g = 0.2$ and $c = 0.7$.

Weighting the sings of life

When using the signs of life in determining residency the problem arises about their reliability: some signs are more reliable than others. Hence it makes sense to weight them. In future we will use three different ways to weight the signs of life. The simple sum of signs of life is defined when all weights in (3) are equal to 1.

The next set of weights used are ratio weights. For defining the ratio weights we use for each year k the sets K_k and N_k of confident residents and confident non-residents

$$K_k = j|R(j,k) = 1, \quad N_k = j|R(j,k) = 0.$$

The sets K_k and N_k are subsets of enlarged population having sizes n_K, and n_N correspondingly, but they do not cover it ($n_K + n_N < N$).

The **ratio weight** of a SOL E_i is the ratio of average value of frequencies among confident residents and average value of frequencies among confident non-residents:

$$b_i = (\frac{n_N}{n_K} \sum_{j \in K_k} E_i(j,k)) / \sum_{j \in N_k} E_i(j,k), \quad i = 1, 2, \ldots, m. \tag{10}$$

To make the weighted sum closer to the simple sum, the values of weights of each SOL are **normalised** using the coefficient T (ratio of average of simple sum and ratio-weighted sum of given SOL):

$$T = (\sum_{j=1}^{N} \sum_{i=1}^{m} E_i(j,k)) / \sum_{j=1}^{N} \sum_{i=1}^{m} b_i E_i(j,k)), \tag{11}$$

$$b_i := Tb_i, \quad i = 1, \ldots, m. \tag{12}$$

The most useful weights were **logarithmic weights.** To reduce the variability of the sum of signs of life, the logarithms of ratio weights b_i in (10) were calculated and normalised using similar coefficient T in (11):

$$q_i = ln(b_i), \quad q_i := Tq_i, \quad i = 1, \ldots, m. \tag{13}$$

The weights q_i are called **logarithmic weights**. In all calculations the sum of signs of life (3) will be used with the logarithmic weights q_i.

Distribution of the sum of signs of life

Sum of signs of life is a random variable depending on person and also on year. Empirical data show that the distribution of sum of signs (3) can be approximated with mixture of two distributions: constant distribution with the only value 0 and a normal distribution. This situation can be explained by the fact that the population consists of two subpopulations: residents and non-residents. The simple sum of values of SOL in the case of residents

follows the central limit theorem and hence can be approximated by normal distribution.

To find the best approximation the following parameters should be estimated: the weight (probabilities) of mixture components and the mean and standard deviation of the normal component. As the signs of life have been measured in Estonia during four years 2012 – 2015 with a small change in methodology in the last year, we can use the empirical data for estimating the parameters, see Table 1.

TABLE 1. Parameters of approximation of the mixture the distribution of sum of signs of life

Year	P(const)	P(normal)	Parameters of normal component	
			Mean	Standard deviation
2012	0,1133	0,8867	4,3048	2,0781
2013	0,1068	0,8932	4,3465	2,1110
2014	0,1068	0,8932	4,3741	2,6102
Total	0,1090	0,8910	4,3418	2,2664

If we compare the normal component of the sum of values of SOL with the theoretical normal distribution then we see that we cannot report the good fit using some statistical test as the sample size is rather big. Still, the visual picture shows that the approximation is acceptable, especially in the case of logarithmic weights, see Figure 2.

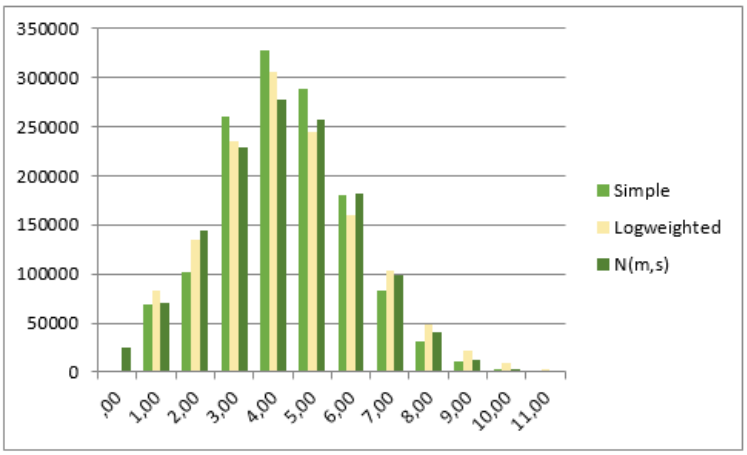

Figure 2. Histograms illustrating the normal component of simple sum of signs, log-weighted sum of signs and the closest normal distribution.

Distribution of $R(k)$ and assessing the residency status

Distribution of $R(k)$ can be considered as a mixture reflexing three groups of population: non-residents having the value $R(k) = 0$, residents having

$R(k) = 1$ and a group of people whose residency status is unclear. The initial distribution of index values of that group can be roughly approximated by the uniform distribution $U(0, 1)$, but it will specified in the process of recalculations. For checking the correctness of decision-making we have to estimate the statistical error arising in estimation of the probability $P(R(k) \geq 0.7)$.

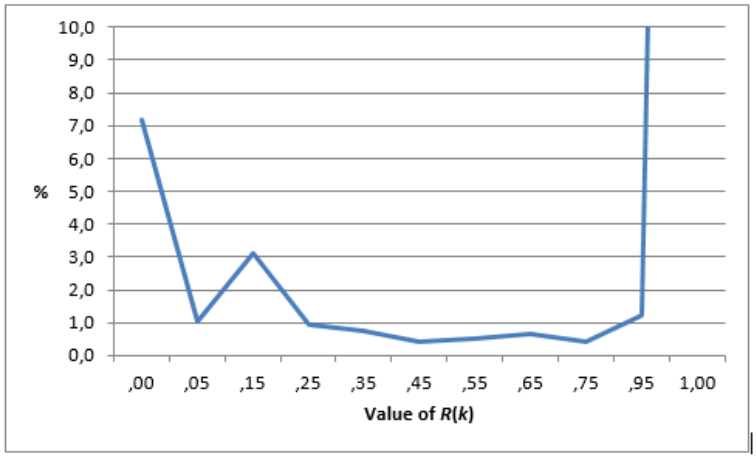

Figure 3. Distribution of index values in area (0, 0.95). Average of years 2013 – 2016.

The empirically estimated weights of the components of $R(k)$, are 0,08; 0,1 and 0,82. From the formulae (6) – (9) it follows that to decide if a person is resident or not the value of his newly computed residency index $R(k)$ must be compared with the threshold $c = 0.7$:

$$\text{if } \begin{cases} 0,8R(k-1) + 0.2X(k-1) \geq 0.7, \text{then the person is resident,} \\ 0,8R(k-1) + 0.2X(k-1) < 0.7, \text{then the person is non-resident.} \end{cases}$$

The possible source of statistical error influencing the adequacy of decision is connected with the recalculation of the index $R(k)$ and estimating the probability

$$P(R(k) \geq 0.7). \tag{14}$$

Common distribution of index and sum of signs of life

It is necessary to estimate the common distribution of variables $R(k-1)$ and $X(k-1)$ to get the distribution of the linear combination

$$0.8R(k-1) + 0.2X(k-1). \tag{15}$$

The first of these variables has the distribution of mixture of two constants, 0 and 1 (with weights 0.12 and 0.85) and uniform distribution U(0, 1). The second variable is the sum of values of SOL having the distribution of mixture

of constant 0 (weight 0.11) and a normal distribution (weight 0.89). We must also take into account that these variables are correlated.

As both distributions are mixtures, we get as a result the 6-component mixture, where the probabilities of components are estimated empirically, see Table 2.

TABLE 2. Common distribution of $R(k-1)$ and $X(k-1)$; empirically estimated weights of components.

$R(k-1)$	$X(k-1)$		Total
	1.Comp.:const=0	2.Comp.:$N(4,34;2,27)$	
1.comp.:const=0	0,0974	0,0311	0,1285
2.comp.:$U(0;1)$	0,0008	0,0147	0,0155
3.comp.:const=1	0,0146	0,8414	0,8560
Total	0,1128	0,8872	1,0000

In the following we analyse the distribution of all components of the mixture and estimate the probability of the event $(X \geq 0.7)$ in all cases.

1. The first component of both variables is the constant 0, also their linear combination (15) equals 0. The weight (probability) of this component of mixture is (by Table 2) 0.0974. (Conditional) probability of the event $(X \geq 0.7)$ is in this case 0, also the probability of estimating the residency (14) is 0.

2. The second component of the variable $R(k)$ is calculated in a linear combination of constant 0 and normal distribution $N(4.34; 2.27)$ multiplied by coefficient 0.2. The component has normal distribution $N(0.868; 0.454)$ and the probability of the event $(X \geq 0.7)$ is in the case 0.645. From Table 2 the weight of this component is 0.0311 and the probability (14) is 0,020.

3. To get the third component of the variable $R(k)$ we have to calculate the linear combination (15) of uniform distribution and constant 0, the result is $U(0; 0.8)$ and the probability of the event $(X \geq 0.7)$ is 0.125. As the weight of this component is 0.0008, then the probability (14) is 0.0001.

4. The fourth component of the variable $R(k)$ is the linear combination of uniform and normal distributions having the weight 0.0147. The probability of the event $(X \geq 0.7)$ can be estimated using a normal distribution $N(\mu; 0.453)$, where $\mu \in [0,868; 1.668]$. As a result we get $P(X \geq 0.7) = 0.868$ and probability (14) is 0.0128.

5. The fifth component of the calculated variable $R(k)$ with the weight 0.0146 is the linear combination (15) of constants 1 and 0 that has the constant value 0.8 and the probability $P(X \geq 0.7) = 1$. As it follows, the probability (14) equals to the weight of the component 0.0146.

6. The sixth component of the calculated variable $R(k)$ with the weight 0.8414 is the linear combination of constant 1 and normal distribution $N(4.34; 2.27)$, where the normal distribution is by assumption truncated

and has no negative values. Hence it follows that $P(X \geq 0.7) = 1$ and the probability (14) equals to the weight 0.8414.

All the steps described are summarised in the following Table 3.

TABLE 3. Calculation of distribution of the residency index $R(k)$.

Component of $R(k-1)$	Component of $X(k-1)$	Weight	Linear combination (15)	$P(X \geq 0.7)$	$P(Res)$	$P(NRes)$
Const=0	Const=0	0,0974	0	0	0	0,0974
Const=0	$N(4,34;2,27)$	0,0311	$N(0,868;0,453)$	0,645	0,020	0,169
$U(0;1)$	Const=0	0,0008	$U(0;0,8)$	0,125	0,0001	0,0007
$U(0;1)$	$N(4,34;2,27)$	0,0147	$U(0;0,8) + N(0,868;0,453)$	0,868	0,0128	0,0019
Const=1	Const=0	0,0146	0,8	1	0,0146	0
Const=1	$N(4,34;2,27)$	0,8414	$N(1,668;0,453), X \geq 0$	1	0,8414	0
Total		1			0,8889	0,1111

The next step is to estimate the standard error of the estimated probability $P(X \geq 0.7)$ in all cases when the decision-making might cause a random error.

TABLE 4. Calculation of standard error and maximal possible number of misclassified persons.

Component	Weight	Population	$P(X \geq 0.7)$	St. error	Width of 95% CI	Misclassified population
$0/N(4,34;2,27)$	0,0311	47 981	0,645	0,002185	0,008738	420
$U(0;1)/0$	0,0008	1 234	0,125	0,009414	0,037655	47
$U(0;1)/N(4,34;2,27)$	0,0147	22 679	0,868	0,002248	0,008991	205
Total						672

Hence, with probability 0.95 the number of persons misclassified due to random error in decision-making process is less than 700, that is less than 0.05 % of population size.

Acknowledgements

This work was supported by institutional research funding IUT34-5 of the Estonian Ministry of Education and Research. The authors thank the referee for valuable comments and suggestions.

References

[1] *Data*. http://www.stat.ee/

[2] E. Maasing, *Eesti alaliste elanike määratlemine registripõhises loenduses*. (Estonian) http://dspace.utlib.ee/dspace;/handle/10062/47557 (2015)

[3] E. Maasing, *First results in determining permanent residency status in register-based census*. banocoss2015/Presentations?preview=#!/preview/149296295/170626623/Maasing_Abstract.pdf (2015)

[4] E.-M. Tiit, *Estimated undercoverage of census 2011*, Quart. Bull. Statist. Estonia **4** (2012), 110–119.

[5] E.-M. Tiit, *Residence testing using registers – conceptual and methodological problems.* https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=#!/preview/149296295/170626640/Tiit_Abstract.pdf (2015)
[6] E.-M. Tiit and, E. Maasing, *Residency index and its usage in population statistics*, Quart. Bull. Statist. Estonia **3** (2016), 41–60.
[7] E.-M. Tiit, K. Meres, and M. Vähi, *Estimation of census population of census 2011*, Quart. Bull. Statist. Estonia **3** (2012), 79–108.
[8] L.-C. Zheng and J. Dunne, *Census like population size estimation based on administrative data.* https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=/149296295/172987273/CensuslikePopulationSize.pdf

Statistics Estonia, Tatari 51, 10134 Tallinn, Estonia
E-mail address: ethel.maasing@stat.ee

Institute of Mathematics and Statistics, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia; Statistics Estonia, Tatari 51, 10134 Tallinn, Estonia
E-mail address: enemargit.tiit@stat.ee

Institute of Mathematics and Statistics, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia
E-mail address: mare.vahi@ut.ee