RESIDENCY INDEX AND ITS APPLICATIONS IN CENSUSES AND POPULATION STATISTICS

Ene-Margit Tiit, Ethel Maasing
Statistics Estonia

How to determine the accurate population figure in the context of constantly increasing mobility of people when immigration and emigration is increasingly difficult to monitor and people are not exactly pedantic about registering their place of residence? Statistics Estonia has found a solution in the form of the residency index.

The issue of accurate population figures

Today all developed countries struggle with the accurate determination or estimation of the population figure. People have become very mobile and they are difficult to get hold of and enumerate. There is also an increasing number of persons who prefer to refrain from disclosing their details to the state for some reason. Immigration – both legal and illegal – only complicates the situation. While traditionally the population figure was determined in a census, even censuses do not produce accurate results nowadays – a part of the population inevitably remains beyond the reach of enumeration, irrespective of the survey method used. A census under-coverage of 2–3% is fairly common in this century, which means at least 10 million un-enumerated persons in the European Union as a whole. Prior establishment of the population figure is very important when a register-based population census is planned.

However, the need for accurate data increases with economic development and updated data is required as frequently as possible. The European Union is planning to cut down the previous 10-year census cycle by a half, or even down to 2–3 years for some characteristics. Collecting and analysing data to make various conclusions and decisions is becoming increasingly relevant. In this, population indicators and especially an accurate population figure constitute the underlying information for all the other indicators, helping to understand phenomena and processes.

Estonian PHC 2011 and census coverage

The latest population and housing census in Estonia was conducted in 2011, with the census moment on the last day of the year. The goal was to achieve as high coverage as possible by enumerating almost all people subject to enumeration, i.e. the permanent residents of Estonia. The following methods were used for this purpose:

- *Respondents were offered an opportunity to fill out the census questionnaire online. Nearly 2/3 of the persons subject to enumeration used this opportunity during the specified period of slightly over a month. It is likely that this increased the motivation for participation among young people, particularly among students.*
- *An enumerator visited the persons and households who did not fill out the census questionnaire online. The period allocated for the census interviews was relatively long (over one month) and during this time enumerators made repeated visits to dwellings where they were unable to contact residents in earlier attempts. Other ways of initiating contact were used as well (such as written messages, phone calls, help from neighbours).*

However under-coverage could not be avoided (see Map, p. 42). A more extensive discussion of the subject matter can be found in the Quarterly Bulletin of Statistics Estonia, issue 4/2012. (Tiit 2012).

When it comes to census over-coverage, the classical reason – repeated enumeration of same persons – was certainly not relevant in PHC 2011 because persons were identified using personal identification codes. The possibility of self-enumeration from abroad was also excluded by asking the respondents to enter the address of their usual residence at the start and excluding all persons whose place of usual residence was not in Estonia from the group of persons subject to enumeration. Of course, the possibility that household members provided incorrect information on the dwellings of some persons cannot be excluded. However, it should be recalled in this context that all censuses are based on the premise that people are honest in their answers. Census data are traditionally based on the respondents' own words and there is no established practice of double-checking such statements (except in the case of obvious errors and discrepancies).

Reducing census under-coverage

Figure 1 (p. 42) shows the differences between different population figure estimates. As soon as the results of the census were established, work began to find a methodology that would increase the accuracy of population estimates, thereby reducing under-coverage, i.e. identifying the persons who were not enumerated and adding them to population records. The available administrative registers of Estonia were used to assess the probability of persons who are entered in the Population Register as residents of Estonia but who were not enumerated as such in the census actually residing in Estonia. Persons who were registered as residents but not enumerated constituted nearly 5% of the population. Several parallel methods (incl. logistic and linear regression analysis and expert assessment) were used for assessment (Tiit 2012; Tiit et al. 2012; Tiit 2015a). As persons of different ages are active in different registers, a series of models was developed to differentiate between residents and non-residents of different ages. The probability of a decision error was also estimated in all cases and it did not exceed 5% (compared to the population described with the model). As a result, the use of 12 registers and sub-registers made it possible to identify 30,000 persons who were very likely permanent residents of Estonia at the moment of the census but were not enumerated (see Figure 2, p. 43).

The population statistics of the subsequent years was produced using the traditional method, based on the estimated population figure of 2012, adjusted for under-coverage. This was done by adding births and registered immigration to the population figure of the preceding year and subtracting deaths and registered emigration. Even though the calculation was person-based and the registration of births and deaths is accurate in Estonia, the resulting population figure was still not accurate enough. This was mainly due to unregistered emigration, combined with unregistered immigration (mostly return migration). Some census data, which were previously accepted as accurate without further checks, also needed further clarification – the potential over-coverage due to enumeration of non-residents was not checked nor taken into account.

Population estimates between censuses. Signs of life

Next we needed a model which would make it possible to use registers for assessing whether a person is a resident at any time, not only during the period immediately following a census when enumerated persons can be used as a control group for fine-tuning the model. This problem was solved by Ethel Maasing in her Master's thesis, in which she analysed all persons, including those who had officially left Estonia, had an Estonian personal identification code and were registered in the Estonian Population Register on 1 January 2015 (approximately 1.5 million persons in total) (Maasing 2015a,b). Several statistical methods (linear and logistic regression analysis, discriminant and cluster analysis) were used to differentiate between residents and non-residents.

The main concept in the work of Maasing, as well as in the works of Ene-Margit Tiit and others (Tiit 2012; Tiit et al. 2012; Tiit 2015a), was 'register activity', which was later renamed by the author as 'sign of life', borrowing from Zheng and Dunne (Zheng and Dunne 2015). A sign of life means some kind of activity in a register in relation to a person at least once in a year. Only the registers which reflect residence in Estonia are considered. For instance, Estonian identity documents can also be changed while abroad and a person does not have to come to Estonia for that; however, a corresponding sign of life was not registered for persons who changed their documents abroad.

In terms of statistics, a sign of life is a binary characteristic, which is contingent on three arguments – person, register and year –, and has a value of 0 if the person in question has not been active in the particular register during the year of observation, or a value of 1 when the person was active in that register at least once during the year. For instance, a person can be assigned a sign of life if he or she visited a family physician at least once or received support from the local government on at least one occasion or was a student at an Estonian educational institution. Using the signs of life and creating different models for major sex and age groups, Maasing succeeded in producing fairly good population estimates, with the best model underestimating the population figure only by a couple of percentages.

The needs of population statistics required a methodology which would facilitate annual estimations of the population figure and composition with sufficient accuracy while also providing a means for calculating external migration. In doing so, it is reasonable to take into account a person's status in previous years. In order to facilitate annual estimation, it made sense to develop a common and fairly flexible methodology for all sex and age groups, which would support ongoing use of additional information (additional registers) as it becomes available.

Residency index

This problem was solved by defining a residency index as an indicator of a particular person's likelihood of being a resident in a given year (Tiit 2015b). The value of the index is calculated for all persons who belong to the 'extended total population', i.e. either the population of the latest census (PHC 2011) or have been registered in the Estonian Population Register (in the period 2012–2016). The extended total population is updated annually based on information obtained from the Estonian Population Register. The persons included in the extended total population can be residents of Estonia or of a foreign country or they can have no residence record at all; furthermore these persons can be listed in the 'passive' section of the Population Register. Consequently, the index is calculated for more than 1.5 million persons in total. This makes it possible to establish the residency of return migrants as well, assuming they have an Estonian personal identification code. Any other non-registered immigrants are generally not detectable by the residency index but the number of such persons (if there are any) is currently low in Estonia.

The value of the residency index changes between 0 and 1. The higher the value of the index, the higher the probability of the person being a resident of Estonia. If a person's residency index has a value of 0, that person is certainly not a resident. A residency index with the value of 1 indicates that the person is certainly a resident. If the value of the index is somewhere in between, then a decision is made based on threshold c: the persons whose residency index is above or equal to the threshold are considered residents, while the persons with an index value below the threshold are considered non-residents.

The index is calculated annually for all persons who were included in the extended total population and were alive at the start of the year; the calculation is based on all available administrative registers (and their independent sub-registers) and on identification of all signs of life established in the preceding year for all persons. For instance, in 2015, the identification of signs of life was based on the Estonian Education Information System, the State Pension Insurance Register, the Health Insurance Information System, etc. The maximum number of signs of life was 27 in 2015, but in the case of a considerable number of persons no signs of life were detected (see Figure 3, p. 45).

Calculation of the residency index

Figure 3, which presents a mix of two distributions, characterises the aggregate of signs of life as a relatively good indicator. One component of the mix is a constant characteristic with the value of 0; the other component has a distribution similar to the normal distribution, with a slight asymmetry (the right-sided tail extends a little further) and a mean value of over 4. Presumably, the right-hand part describes Estonian residents and the left-hand part non-residents, while the status of persons with only one sign of life is not quite clear at a first glance. However, signs of life alone are not reliable enough as bases for making the decision. It must be taken into

account that it is quite possible that some Estonian residents do not show any signs of life in the course of a year, while some non-residents can have several signs of life identified in the registers.

Additional information on a person's residency can be obtained by looking at the person's status in previous years, because even today people are not changing their places of residence too frequently. Consequently, the residency index $R(j,k)$ of a person number j ($j = 1,\dots,N$) in the extended total population (with volume N) in year k is defined as follows:

$$R(j,k) = d \cdot R\Big(j,\big(k-1\big)\Big) + g \cdot X(j,(k-1)), \qquad (1)$$

where $R(j,(k-1))$ is the person's residency index in the previous year and $X(j,(k-1))$ is the weighted sum of the person's signs of life accumulated in the previous year,

$$X(j,k) = \sum_{i=1}^{m} a_i\, E_i(j,k), \qquad (2)$$

where $E_1(j,k), \dots, E_m(j,k)$ indicates the signs of life accumulated by person j in year k:

$$E_i(j,k) = \begin{cases} 1, if\ person\ j\ is\ active\ in\ register\ i\ in\ year\ k \\ 0, otherwise \end{cases}$$

and m is the maximum possible number of signs of life in that year ($i = (1,\dots,m)$).

In Formula (2), a sign of life number i has a weight of a_i. In the simplest case, all weights a_i equal 1; in this case $X(j,k)$ in Formula (2) represents a simple sum of signs of life, the distribution of which in 2015 is shown in Figure 3 (p. 45).

Establishing the parameters of the residency index

The values of multipliers d (stability rate) and g (signs of life rate) are selected based on rationality of decisions. A higher d value indicates greater impact of the past status on a person's residency. A higher g value indicates greater impact of signs of life. It is reasonable to set parameters d and g and threshold c at a level where a definite resident (i.e. a resident with a residency index value of 1) does not drop out from the category of residents if there are no signs of life in the course of one year, but loses the residency status if there are no signs of life for two years in a row. It means that the following formula should apply:

$$d^2 < c \leq d.$$

It is also reasonable to assume that a single sign of life is not sufficient for a definite non-resident (i.e. a person with residency index value of 0) to become a resident, even if the signs of life appear in several years, but it could be possible if the definite non-resident has several signs of life.

If a previously definite non-resident has one sign of life every year, his or her index in the first year would be g, then it would be dg + g in the second year and in year s it would be:

$$g(1 + d + \dots + d^{s-1}) = g(d^s - 1)/(d - 1). \qquad (3)$$

A person can be classified as a resident after s years if the value of expression (3) exceeds threshold c.

Based on initial assessment, the suitable values of the parameters were $d = 0.8$, $g = 0.2\ and\ c = 0.7$.

This means that a person loses residency if he or she does not have any signs of life in two years. If a definite non-resident has one sign of life for six years in a row, he or she becomes a resident. The status of a resident can be assigned earlier in the case of a higher number of signs of life. The adequacy of selected parameters was confirmed through statistical verification of data, incl. analysis of migration data and comparison with survey data from 2012–2015.

According to the definition, persons are classified as residents if their index value is at least equal to the threshold, while all persons whose index is below the threshold are excluded from the category of residents. For the purpose of future use, any index values higher than 1 are truncated to 1, which means that it is actually irrelevant in terms of residency decision whether a person has a high or very high number of signs of life.

All conventional population events are fully taken into account in the calculation of the residency index: if person j is born or registers immigration in the year $k-1$ *, the person's index will have a value of* $R(j,k)=1$*; if the person officially leaves the country, the index will have a value of* $R(j,k)=0$*, but the person will be kept in the extended total population as a person who could potentially regain the status of a resident in the future. Upon death the person will be excluded from the extended total population.*

Formula (1) includes as special cases the previously described methods of population accounting. If d = 1 and g = 0 then we use traditional population accounting where signs of life are not taken into account and the population figure of the preceding year is only adjusted for population events. If d = 0 then we use the model-based population accounting as presented in the Master's thesis of Maasing (Maasing 2015a), without considering the person's status in the preceding year and with weights a_i *assigned to signs of life according to the selected statistical procedure.*

Weighting of signs of life

A disadvantage of formula (1) in the case of simple sums stems from the fact that not all signs of life are equivalent in differentiating between residents and non-residents. For instance, if a person is permanently residing in an Estonian care home, that person is definitely a resident; however, an Estonian driving licence can also be issued to a person who is usually residing abroad and visits Estonia only to renew the driver's licence. Therefore, it is reasonable to assign the signs of life in formula (2) such weights a_i *which correspond to their reliability in determination of residency.*

First, we define ***relative weights*** b_i*. These are calculated from the data of the preceding year. For each sign of life, its average frequency among definite residents and definite non-residents is calculated. Assume that the number of definite residents in year k is* K_k *and the number of definite non-residents is* N_k*:*

$$K_k=\{j: R(j,k)=1\}, \qquad N_k=\{j: R(j,k)=0\}$$

Then we calculate the frequencies of each sign of life E_i *in both sets and define relative weights* b_i *as the following ratio of frequencies:*

$$b_i=\left(\sum_{j\in \boldsymbol{K}_k} E_i(j,k)\right)/\left(\sum_{j\in \boldsymbol{N}_k} E_i(j,k)\right),\ i=1,2,\dots,m. \qquad (4)$$

Relative weights are not calculated for those signs of life that never appear in the case of definite non-residents, i.e. where the denominator of expression (4) equals zero. Such signs of life can be considered to be very reliable and their relative weight is defined through the maximum of relative weights of the remaining signs of life:

$$b_I=max_i\, b_i+1,$$

where I is a sign of life that never appears in the case of definite non-residents.

In order for the mean parameter values to correspond to the indexes calculated with weighted signs of life, we normalise the relative weights, i.e. multiply all relative weights calculated according to formula (4) by factor T, which is found by dividing the simple sums of signs of life for all persons with the sums of weighted sums of signs of life for all persons:

$$T=\left(\sum_{j=1}^{N}\sum_{i=1}^{m} E_i(j,k)\right)/\sum_{j=1}^{N}\sum_{i=1}^{m} b_i E_i(j,k) \qquad (5)$$

$$b_i \coloneqq Tb_i, \qquad i=1,\dots,m.$$

A higher relative weight indicates greater reliability of a sign of life in residency determination. However, it turned out that the relative weights were quite different (with some difference factors in multiples of ten), which created the possibility of the index becoming unstable (through amplification of any differences between years). To avoid this, it was decided to use logarithmic weights, which were obtained by finding the logarithms of relative weights and normalising them.

Calculation of ***logarithmic weights*** q_i

$$q_i = ln(b_i),\ q_i := Tq_i, i = 1, \dots, m. \qquad (6)$$

Formula (6) features the normalising factor T, which is calculated in formula (5) using logarithmic weights q_i instead of relative weights b_i.

Figure 4 (p. 47) indicates that the weights vary somewhat year to year. This could be caused by random fluctuations or by changes in registry policies. This explains why it is reasonable to calculate new weights for each year. Figure 5 shows that a logarithmic weight can also have a negative value (if the relative weight is lower than 1). This indicates a sign of life which is unusually more common for non-residents than residents. This is the case with residence permits, the holders of which are quite likely to leave Estonia in the near future. This also creates the possibility that the calculation results in a negative value of the residency index. In this case the truncation rule should be updated and any index with a negative calculated value should be considered equal with 0.

Using the residency index for population figure calculation

In order to test the methodology based on the residency index, several methods have been used for population figure calculation since 2012, including differently weighted residency indexes (see Figure 5, p. 48).

The results indicate that the difference between population estimates using differently weighted signs of life was half a percentage of population in three years. The closest result to the published population figure was the estimate calculated using signs of life with logarithmic weights. All index-based population estimates indicate a relatively large population size in 2013: this is the year following the census (the population size according to the census was the basis for the 2012 estimate) when loss of residency was prevented by the criterion of two years, arising from the definition, i.e. the selection of parameter d. It is likely that this estimate includes persons who were counted as residents during the census but were excluded from the population in the course of subsequent years.

From 2016, Statistics Estonia uses the residency index methodology for estimating the population size and calculates the index using signs of life with logarithmic weights. The moment of transition resulted in an increase in population by 0.3% compared to the previous estimate (the estimated population size increased). However, the population estimate of Statistics Estonia, based on the residency index, is over 2.5% smaller than the population figure according to the Population Register. This difference is probably caused by an assumed over-coverage of the Population Register due to persons who have left the country without registration (see Figure 6, p. 49).

Using the residency index for external migration calculation

It is only natural that migration events are consistent with index-based population accounts. External migration generally causes a change in a person's residency status. This fact was also used to test the index-based methodology, incl. the adequacy of the determinant parameters of the index. As expected, the index-based estimate of external migration was somewhat higher than registered external migration, as it also included unregistered external migration. Index-based commuting (alternating immigration and emigration) was also compared to the volume of commuting determined by a survey in the same period and there was a good degree of consistency between the datasets. Unregistered external migration is still of the same order of magnitude as registered external migration, but unregistered immigration, including return

migration, has appeared as a new trend which includes Estonian residents who have returned without having previously registered their departure.

The residency index uses simple and logical definitions of migration acts:

Person j is immigrating *if his or her residency index is 0 in year* $k-1$ *and becomes 1 in year k, i.e. the following equalities apply:*

$$R(j,k-1)=0 \; and \; R(j,k)=1$$

and it is not a case of birth.

Person j is emigrating *if his or her residency index, which was 1 in year* $k-1$ *, becomes 0 in year k, i.e. when*

$$R(j,k-1)=1 \; and \; R(j,k)=0$$

and it is not a case of death.

In the case of immigration, it is also important to establish the person's place of residence in Estonia. If the person has not officially registered the act of migration, their previous place of residence (from the Population Register or census data) can be used as the place of residence. However, if the person does not have an official registered place of residence in Estonia at the start nor at the end of the migration year, the person will remain on a 'pending list' for one year. It means that the person cannot be classified as a resident (counted among permanent residents) in that year. If the person's residency index value is again 1 the next year, i.e. $R(j,k+1)=1$ *and the person still has no registered place of residence, the person is counted as a permanent resident with unknown place of residence. The methodology based on the residency index does not require knowledge of the previous country of residence of immigrants: the country of origin can be unknown.*

The index-based calculation of external migration was implemented in 2016. As per usual, the change in methodology resulted in a leap in statistical indicators: immigration increased by more than four times and emigration by more than two times compared to the mean value of the preceding three years (Figure 7, p. 50). The previously negative migration balance was now moderately positive, however, this was not a result of the methodology change but an actual change in the trends of external migration. However, all these initially unexpected changes have a very logical explanation. Having used as the basis the population size at the start of 2012, when census data were verified based on a model, the migration statistics of subsequent years – 2012, 2013 and 2014 – only included registered migration. The transition to index-based statistics at the start of 2016 essentially meant that all unregistered migration events of the preceding years were now taken into account as well: this includes unregistered departures and returns of those who had left without registering their departure (registration of the latter was not even possible). Furthermore, the new methodology also included immigration events without a specified country of origin, which were previously not included in migration accounts. Such incoming persons (mostly from the European Union) constituted more than half of all persons who were registered in the past year as immigrants, but in fact they had come to Estonia at various points in time in the past three years. Considering the fact that the results cover the migration process of the previous four years and comparing registered and unregistered migration, we assumed that unregistered emigration has been uniform in all the years while unregistered immigration has increased (due to the law of arithmetic sequences). The picture revealed by this model was much closer to the one that could have been expected (see Figure 8, p. 50), with immigration and emigration figures of 2012–2015 modified so that all aggregate indicators (as at 1 January 2016) remained the same.

A considerable increase in migration flows can also be anticipated in the future in connection with additional accounting of unregistered migration compared to previous statistics, which only included registered migration.

Using the residency index for internal migration calculation

As the persons who are included in the category of residents based on the index are assumed to have a place of residence in Estonia, the methodology based on the index makes it possible to calculate the number of registered residents for every local government, city and county, whereas the place of residence is preferably taken from the Population Register if the person has an officially registered place of residence. If this information is not available, the place of residence established in the census is used or the place of residence of the mother is used for children; if this information is also unavailable, the person's place of residence is entered as 'unknown'. There were more than 1,500 residents in Estonia (0.12%) at the start of 2016 whose county of residence was unknown.

The index-based calculation of internal migration of residents is similar to the external migration calculation, but the estimation of unregistered migration is generally not possible in the case of internal migration, because most registers do not provide more accurate information on places of residence than the Population Register, which shows the legally registered places of residence.

A resident is considered to have left a county, city, local government unit or city without municipal status when he or she was a resident in year $k-1$ *and his or her place of residence was located in that administrative unit, but the person is no longer a resident in year k or the person continues to be a resident but his or her place of residence is in a different county, city, local government unit or city without municipal status.*

A resident is considered to have arrived in a county, city, local government unit or city without municipal status when he or she was not a resident in year $k-1$ *but has acquired the status of a resident by the year k with a place of residence in that administrative unit, or the person was a resident in both year* $k-1$ *and year k but had a place of residence elsewhere in year* $k-1$ *but has moved that particular county, city, local government unit or city without municipal status by the year k.*

Arrival in a county, city, local government unit or city without municipal status as a result of birth, or departure as a result of death, is not included in the internal migration accounts. However, both births and deaths can be combined with internal migration (when the mother's place of residence at the time of birth differs from the place of residence at the end of the year or when the place of residence at the time of death differs from that person's place of residence at the start of the year). Internal migration can be simultaneous with an external migration event when the person also crosses the state border.

We would like to thank Alis Tammur and Koit Meres for their methodological work and calculations on migration.