Papers on Anthropology XXVI/1, 2017, pp. 72–87

INDEXES IN DEMOGRAPHIC STATISTICS: A METHODOLOGY USING NONSTANDARD INFORMATION FOR SOLVING CRITICAL PROBLEMS

Ene-Margit Tiit, Mare Vähi

Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia

ABSTRACT

A new methodology for solving different problems in population statistics (estimating under- and overcoverage of census, determining the population size and residency of persons, finding the partners and placing persons into living rooms) is presented. In all cases, so-called signs of life, demonstrating the activity of persons in different registers are taken as arguments or explanatory variables for models. The weights of models are calculated using training data, when the models are in use sequentially; then every year the weights are recalculated using the data of the previous year.

Keywords: *residency; population size; registers; partnership; family type*

INTRODUCTION

New values in society, for example, very highly coveted privacy, openness and mobility, lead to new traits in people's behaviour that make the tasks of demographic statisticians more and more complicated. Today, many persons do not like to be counted, also they do not like to register their migration from one country to another. Unregistered cohabitation instead of marriage is also very common. This means that nowadays traditional methods of demography based on interviews and registration of events do not work perfectly. Instead of counting and calculating, it is necessary to use assessing and statistical estimation. But here also new problems arise: in demography, the methods of classical statistics usually give rather poor results. Usually, in demographic tasks, the sample sizes are enormously large and the traditional rules of statistics do not

https://doi.org/10.12697/poa.2017.26.1.06

fit. Often the variables that can be used as explanatories have non-traditional form and distribution, more and more often it is challenging to use big data or combinations of big data and register data, but few good methodologies exist in this area.

Nevertheless, demographers and statisticians have to calculate population size, the sizes of sex-age groups, internal and international migration flows, etc. All of this is necessary for estimating the current status of population, but also for making population forecasts. Also, there is an obligation to report demographic data to international organisations (EUROSTAT, UN).

When speaking about the existing demographic data, the situation is somewhat controversial: the census data have lost part of their credibility due to high mobility of the population as well as the refusal to collaborate, which causes under-coverage of census data. At the same time, the variety and amount of data on population is rapidly increasing: censuses contain much more information than before, there is a number of administrative and other registers collecting data on different groups of population, and big surveys containing rich material have been conducted. There also exists another kind of information – the so-called big data – that can also be identified and linked with population data. But still it is not clear how to use these different types of information for making demographic calculations.

In the following sections, four demographic tasks that emerged during the preparation of the register-based census and their possible non-traditional solutions are regarded. Two of the tasks have been solved in Estonia using nonstandard data and methodology. One is in process, and for the last one, only the possible methodology for a solution is proposed.

TASK ONE: ESTIMATING CENSUS UNDER-COVERAGE

Estimating census under-coverage in Estonia using signs of life (SOL)

First, it is shown how nonstandard information was used for solving one of the most basic tasks when using the results of a traditional census – estimating under-coverage and calculating the actual population size.

In Estonia, the need to estimate the actual population size arose after the 2011 Population and Housing Census [1]. After the census period had finished, the census team received a multitude of messages and phone-calls from people who were not enumerated for different reasons, in spite of the possibility of self-enumeration on the internet and also face-to-face interviews during quite a long period (92 days in total). Comparing three population sizes – the census

population size, the population size calculated currently by Statistics Estonia and the number of respondents in the Estonian Population Register – the differences were about 2–5%. From the supplementary information, the census team saw that there was some under-coverage in the census results.

The first step was to improve the census data using different registers available in Estonia. Here, the following concepts and decision process were used [2, 3]:

1. The persons who were listed in the Population Register as Estonian residents but who were not enumerated were considered the research population (size about 60,000 persons that is 5% of the population). The task was to divide them into two groups: residents and non-residents.
2. Two test groups were formed: confident residents, who were enumerated as Estonian residents and also had the status of an Estonian resident in the Population Register, and confident non-residents, who were not enumerated in Estonia and were not Estonian residents in the Population Register.
3. For supplementary information, a number of administrative registers were chosen, and in all the registers for each person from the research population and test populations their activity during the census year was checked.
4. For each person the signs of life (SOL) were determined in the following way: if person j had been active in register i during the census year, then he/she got a sign of life; that is binary variable $E(i)$ had value $E(i,j) = 1$, in all other cases $E(i,j) = 0$.
5. Using the test groups the forecasting models (logarithmic and linear regression) were created.
6. It turned out that the distribution of SOLs was quite different in different sex-age groups. In order to get better results, the whole population (including both research and test populations) was divided into twelve sex-age groups and in each of these groups the best model was selected.
7. The models were used for the research population in order to identify who was a resident (the group of under-coverage) and who was not a resident (error in the Population Register). Also, the inclusion and exclusion errors were calculated.
8. From all of the models, the model minimizing both errors was selected for practical use.
9. The common decision rule had total inclusion and exclusion errors less than 5% (from the research population, not more than 3,000 persons).

10. The estimated under-coverage was added to the census population, and in all demographic calculations the revised population size was used. As for the persons in the Population Register, all the main demographic indicators (age, sex, residence, citizenship, legal family status, etc.) have been fixed, so improvements were also made in all subgroups of the population.

TASK TWO: ESTIMATING POPULATION SIZE

Current estimation of population size

The following task was the current estimation of the number of residents in the country in the years following the census. Here, the problem is that emigration is not always registered and also the returners (who had not registered their leaving) do not register their immigration. Hence, the calculated population size that was more or less exactly estimated immediately after the census will in time decline from its actual value.

The problems of estimating population size and estimating international migration are connected. When there exists a methodology for calculating the population size for each year, it is also possible to calculate (estimate) net migration, as the exact data for natural increase are known. From here also follows the possibility to assess non-registered migration that has been completely unknown so far. Knowledge of the exact population size is also very important in the preparation process for the register-based census [4].

The problem was solved in Estonia by Ethel Maasing and Mare Vähi in the following way. As a basis for models for estimating the actual population size, signs of life (SOL) from about 20 registers were used again. It was reasonable to make the calculations yearly and to use for calculations for year k the population calculated for year k–1. The following problem was to define the research population that would include all the possible residents. In the case of people who belonged to the set of residents last year, it is necessary to check whether they have not left during the year. It is more complicated to find and check the possible immigrants. It is reasonable to cerebrate here in the following way: everybody who enters the country for the first time must register themselves (in Estonia, the rate of illegal immigration is almost zero). The non-registered immigrants are people who have previously been Estonian residents or have lived in the country temporarily. Hence, there must be a record of each such person in the Population Register – either in the part of non-residents or possibly in the archive. In such a way, the research population, including the

population of "potential residents" (e.g. the set of all persons having an Estonian ID and being recorded in any of the Estonian registers) was created.

The test populations **N** and **R** were formed using both the census and the Population Register data: confident residents (population **R**) were persons enumerated as residents and belonging to the Population Register; confident non-residents (**N**) were the persons not enumerated (and not estimated as under-coverage) and not belonging to the Population Register. In the future, when the calculations have been made for several years already, the population of the previous year will be used as test groups. The sizes of the test groups are $n(\mathbf{R})$ and $n(\mathbf{N})$. The situation is somewhat different from the earlier case: both test groups belong as a part to the research group, but this situation has no influence on results. Similarly to the task of estimating under-coverage, the whole population was divided into sex-age groups.

The decision model was built using several methods of multivariate statistics: linear and logistic regression, discriminant analysis and taxonomy, using SOLs as explanatory variables. In all the cases, the errors were estimated. The estimated population of residents was compared with the current statistics, for which the traditional methodology of calculating current population was used. It turned out that all models produced somewhat under-covered results. Still, when using the best model, the result was, in general, satisfactory – the under-coverage was about 1–2%. [5, 6]. But in order to use the methodology in demographic calculations that will be published as time series, the accuracy should be better.

Using the sum of SOLs

When analysing the connection of SOLs with the research group, it became evident that one reason of erroneous decisions was the heterogeneity of the research population, especially existence of small subgroups who had SOLs that differed from the rest of the population. For instance, one group of almost sure residents were people living in a nursing home. But as the number of such persons was quite small, the SOL "living in a nursing home" was not included in the model (or its influence was very small compared with other explanatory variables). So, the people having SOLs that occurred rarely were not included in the set of residents by the model. In a similar way, many conscripts were not included, as the SOL "conscript" emerged seldom (and usually the conscripts had no other SOLs). From here, the idea emerged to build the model in such a way that in all cases all SOLs (the total number of SOLs is m) would be present. In this case, it was not necessary to divide the population into subgroups that made the model-building easier and less time-consuming.

As the first step, the simple sum of SOLs was used. A similar idea has also been mentioned in [7]. It means that, for each person *j* from the research population, sum X_j was calculated in the following way:

$$X_j = \sum_{i=1}^{m} E(i,j), \tag{1}$$

where $E(i,j)$ means the value of SOL *i* in the case of person *j*. Variable X_j is a simple sum of SOLs. The distribution of variable X_j in the research group was clearly a mixture of two distributions: one close to Normal distribution, the other close to constant distribution of value 0. It was logical to rate that the first group describes residents, the second group – non-residents. For making decisions about the status of a person *j*, it is necessary to fix threshold *c* for X_j, so that if $X_j \geq c$, then person *j* belongs to the set of residents and in the opposite case – non-residents. For assessing the value of threshold *c*, the empirical data of the test population can be used.

Weighting SOLs

In using the simple sum of SOLs, one additional problem became evident. It is clear that all SOLs do not have the same weight in the sense of information about residency status. Several SOLs indicate almost surely that a person is a resident, but such SOLs occur rather seldom. Such are both SOLs named earlier as an example – living in a nursery home and being a conscript. Some other SOLs are quite common, e.g. visiting doctors, but in some cases non-residents can also get such SOLs. Hence, it is necessary to weight the SOLs so that the weights are proportional to the impact of SOLs considering the residency status of persons.

To assess this information, the weights were calculated in the following way. Each SOL *i* can, in general, occur in both test populations **R** and **N**. Let us consider that the set of persons having $E(i,j)=1$ consists, in general, of persons from **R** and persons from **N**. Let the number of the first group be $u(i,\mathbf{R})$ and the number of the second group be $u(i,\mathbf{N})$. Then we calculate the weight $q(i)$:

$$q(i) = [u(i,\boldsymbol{R})/n(\boldsymbol{R})]/[u(i,\boldsymbol{N})/n(\boldsymbol{N})] \tag{2}$$

that equals the ratio of average frequencies of SOL *i* in test populations **R** and **N**. It may happen that $q(i,\mathbf{N}) = 0$, in which case the highest weight (among the cases when the ratio is calculable) is ascribed to $q(i)$:

$$q(i) = \max_{1 \le v \le m} q(v).$$

It is logical to use only the SOLs with weights that fulfil the condition $q(i) > 1$. In the opposite case, the SOL i should be excluded from the set of SOLs.

Sometimes it is useful to apply to weights $q(i)$ a monotonic transformation. In our case, the variability of weights $q(i)$ was quite high, and it was reasonable to use logarithms of weights instead of their original values. Then, also the so-called "negative" SOLs with higher impact for non-residents than for residents were useable. In our case, there was one such SOL – residence permission given to persons when they immigrated. These people had high probability of leaving the country the next year. The last step in defining weights was their standardisation using constant K, calculated in the following way:

$$EX = \frac{1}{N} \sum_{j=1}^{N} X_j , \quad EQ = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{m} q_i E(i,j);$$
$$K = \frac{EX}{EQ}. \tag{3}$$

When using weights Kq_i, all theoretical calculations and conclusions made for simple sums are in a more general case, in average, valid.

By using the model with logarithmic weights for assessing residency we were able to improve the result markedly [8].

Stabilisation terms in the model

Nowadays, people are quite mobile, and it happens quite often that a person lives in different countries during a year or lives part of a year abroad and then returns. When assessing residency using the (weighted) sum of SOLs and using the same information for calculating migration, it may happen that if many people are commuting between several countries, the numbers of international migration will markedly increase compared with traditional migration statistics. This demonstrates the fact that model-based migration statistics are more sensitive and precise compared with traditional methods, but in this case, it is not an advantage. Traditionally, there is some time lag before changing residency. Hence, it is logical to use information on past residency status in making decisions about residency in the current year.

From here it follows that it is also useful to add to the model a stabilising term that shows the person's status in the last year. Hence, we created the formula:

$$R(j,k) = dR(j,k-1) + g\sum_{i=1}^{m} q_i\, E(i,j,k-1), \qquad (4)$$

where the values of indexes $R(j,k)$ must be truncated to fulfil the conditions $0 \leq R(j,k) \leq 1$. Person j is assessed to be a resident in year k, if $R(j,k) \geq c$, and a non-resident in the opposite case. The parameters d, g and c should be estimated using empirical data (test population or additional surveys). Also, some logical considerations about the duration of being resident without any SOLs and getting the status of resident by SOLs are counted in calculation of the parameter's values. The initial conditions for parameters are: $0 \leq c, d, g \leq 1$, $d+g=1$.

The estimated values of parameters in Estonia, used in practical calculations, are $d=0.8$; $g=0.2$ and $c=0.7$ [4]. When using such values, a person can on the average be a resident without any SOL for two years, and is then excluded. With only one SOL (of medium impact), a person must wait on the average for five years before getting the status of a resident, but with at least five SOLs, it is possible to get the residency status already the next year. Additionally, it has been requested that the person has a permanent place of residence in the country. All these calculations apply to SOLs with average impact – in particular cases of SOLs with very high or low impact, the numbers of years can change.

The decision errors of the method were assessed statistically by estimating the distribution of index (before truncating) in the research population. It turned out that the distribution was bimodal in the form of a mixture of two components. One of them has a distribution close to the normal distribution (mean about 4, as on average each person from the set **R** has about four SOLs), and the other component is close to a constant distribution of value 0. Near the threshold, the number of points (persons) is quite small. Distribution of points near the threshold was specially analysed, the possible errors of parameters were also estimated. As a result, it turned out that the inclusion and exclusion errors were both about 0.1%. The distribution of parameters was also checked via simulation. The results are also consistent with estimates from surveys [9].

As the errors in model-based calculations were significantly fewer than errors cumulating when using traditional methodology due to incomplete registration of migration, a decision was made to use the model-based population size (number of residents) in demographic calculations. Since 2015, the population size of Estonia has been calculated using the index methodology. The population account is personalised, using encrypted (anonymised) IDs. For each person, a decision has been made about his/her residency status in

the current year, and these calculations will be made for the following years as well [10].

The determination of the set of residents is also important for the register-based census, as in this case, the census population must be defined beforehand, and the registers to be used have, in general, different population sizes. The census population defined by using the model was implemented in Estonia when a pilot census was carried out in 2016.

TASK THREE: ASSESSING THE ACTUAL PLACE OF RESIDENCE

Assessing the actual place of residence

The methodology used successfully in calculation of population size might also be useful for solving some other problems in statistical demography. Here, we will consider the problem connected with false registration of places of residence. In the register-based census, it is impossible to check the correctness of registered places of residence of residents, as it is in the case of face-to-face interviews. It may happen that there are many incorrect addresses, especially when different bonuses exist in different cities and counties, and the registration of places of residence is free, or the registration of a false address is not punishable.

Let us regard the task of defining actual places of residence. The task is different from the task of assessing the residency status, as two populations must be analysed here. One of them is population **P** of all adult residents who do not at the moment live in any institutional household. The second population is the population of dwellings **S**. The potential research population is the set of all pairs (*j,s*), where *j* is an adult person from set **P** of residents and *s* is a dwelling from set **S** of all (liveable) dwellings. Here the research population **P×S** is very large, and for practical work, it is useful to restrict it.

For assessing the actual place of residence of residents, it is necessary to use instead of SOLs signs of dwelling placement (SODs) connecting the persons $j \in \mathbf{P}$ with their (actual) places of residence $s \in \mathbf{S}$. The SODs linking the points from the two populations are binary variables: $D(j,s;i) = 1$, if the SOD i (i=1,2,…,m) links person j with dwelling s and $D(j,s;i) = 0$ vice versa.

The list of SODs might include person's addresses (in different registers) and some so-called big data: information of mobile positioning, information about paying different payments and taxes connected with the residence (taxation for real estate, rent, payments for water supply, electricity, central heating,

etc). SODs might also include places of residence of family members (parents, children, spouse) if these are known.

For practical work, it is reasonable to extract from the set of all pairs only the pairs with at least one positive SOD. This set of pairs will be considered as research population in the future. By using the SODs, it is possible to define and solve several different tasks (connected with each other):

1. To find for each person (adult resident) his/her actual place of residence (and a secondary dwelling, if it exists).
2. To check for each dwelling if it is inhabited or not and find out who live in this dwelling.
3. As the set of people sharing a dwelling form a household, this is also a methodology for the determination of households.

Test populations

For solving the problems by using statistical methodology, it is necessary to have test populations. The test population of fitting pairs (a certain person who actually lives in a certain dwelling) can be formed on the basis of surveys where the information on actual places of residence of people has been obtained (as a result of a face-to-face interview). In forming the test population **R,** it is important to form a pair for each adult person living in a dwelling; hence, one dwelling can be the residence of several persons and form different pairs all belonging to **R**. Besides the set **R** of "positive pairs" (consisting of person *j* and his/her actual dwelling *s*), it is also necessary to form the set **N** of "negative pairs" with several SODs. Here are some options:

1. One option is to form all possible pairs (person and dwelling) from the test population and subtract from them all "positive" pairs, that is population **R**.
2. Another option is to take a sample from the population of all possible pairs using a method of random choice.
3. The third option for creating population **N** is to design a special population using the persons from "positive pairs" and checking if they have SODs connecting them with another dwelling, and in a similar way to find pairs containing dwellings from **R** with "false" inhabitants, who have some connections with the given dwelling.

Assigning to pairs from **R** ("positive pairs") index value $D(j,s;i)=1$ and to pairs from **N** ("negative pairs") value $D(j,s;i)=0$ and using SODs as binary predictive variables for regression models (logistic and linear), it will be possible to build a model for checking the actual place of residence.

The next step is to apply the model created using test populations to all pairs consisting of adult residents *j* and dwellings *s* linked with the person at least with one SOD. As a result, we have for all pairs (*j,s*) from the research population estimated the value of placement index *D*(*j,s*)

$$D(j,s) = \sum_{i=1}^{m} b_i D(j,s;i) \quad (5)$$

where *D*(*j,s;i*) is a SOD connecting person *j* and dwelling *s* and coefficient b_i is a model coefficient estimated in a traditional way.

If the value of *D*(*j,s*) is higher than threshold *c*, it can be concluded that person *j* lives *de facto* in dwelling *s*. If there are two dwellings where a person lives, the dwelling having a higher index is considered to be his/her permanent place of residence and the other (others) his/her secondary dwelling(s). Again, threshold *c* should be estimated by empirical data. This methodology has not been used in practice so far. Some additional methodological steps are suggested:

1) Instead of using standard regression analysis, it is also possible to use weighted coefficients, see (2) and (3).
2) It might be useful to add the stabilising term (see (4)) in calculation of placement index *D*(*j,s*).

TASK FOUR: ASSESSING PARTNERSHIP

Assessing partnership

One serious problem is assessing partnership in the situation when partners are officially living in different dwellings – a situation that is quite common in Estonia due to different benefits offered by the local governments of cities, towns and communities. This task is connected with the task of actual places of residence, but we will see how this task can be solved separately. Solving these tasks separately gives a possibility to check the solutions achieved in different ways. The census team of Statistics Estonia has started to solve this task before the register-based census in 2020, but there are no positive results yet.

In this case, the research population is the set of all pairs (*h, j*) of adult persons belonging to population **P** of residents who do not live at institutions, do not live together and fulfil the following conditions:

1. are free (that is they do not belong to any household as a partner);
2. are of opposite sex;
3. the age difference of persons *h* and *j* is not too big (<19 years).

There might be some additional conditions to restrict the research population. In the first stage, we will consider the pairs where at least one partner is a single parent.

The next step is to fix the list of signs of partnership (SOPs). The SOPs are: registered marriage or cohabitation, common children, common property (real estate, car), common duties, mutual recognition. Some signs of partnership indicating the common dwelling of partners can be useable in solving the problem; some SOPs could also be defined on the basis of big data (mobile positioning). In general, the SOP $P(h,j;i)$ is a binary variable having value 1 if h and j have the linking sign i and 0 in the opposite case. The total number of SOPs defined is m.

For model-building, survey data can be used again as test population. The group of "positive pairs" **R** consists of pairs (h, j) who form (by survey information) an actual couple (or cohabiting pair), but their registered addresses might be different. For all these pairs the partnership index $P(h,j) = 1$.

For defining the "negative" test population of pairs with $P(h,j) = 0$ there are again several possibilities:

1. All the pairs from the test population who do not belong to population **R**.
2. A sample from the population of all possible pairs formed by random choice.
3. A special design of test population **N** of "negative pairs" is the following: find for all persons j and h belonging to "positive pairs" a "false" partner linked with him/her with at least one SOP. All these pairs of "false" partners are "negative" ($P(h,j)=0$) and form the test population **N**.

The following steps are similar to the process described in the solution of the previous tasks. The regression models for partnership index $P(h,j)$ are created using test populations. Then the models are applied to the whole research population and for all pairs having at least one SOP the partnership index $P(h,j)$ is calculated:

$$P(h,j) = \sum_{i=1}^{m} b_i \, P(h,j;i). \tag{6}$$

Persons h and j, satisfying conditions (1–4), are considered as a pair (couple), if $P(h,j) \geq c$. Parameter c can be estimated by empirical data. Again, there exist some additional steps that might be useful:

1) Instead of using standard regression analysis, it is also possible to use weighted coefficients defined by formulae (2) and (3);

2) It might be useful to add the stabilising term (4) in the calculation of the partnership index $P(j,s)$.

INDEX AS A METHODOLOGICAL TOOL FOR SOLVING DEMOGRAPHIC PROBLEMS

Defining "signs" on the basis of different information sources

From the examples proposed above, it follows that the classical multivariate methods might not give the best results in solving demographic problems when the amounts of data are huge (several millions or more) and the data are in some sense unusual (taken from registers or collected as big data not for statistical purposes). This means that assumptions used in traditional statistics are not fulfilled. For solving such problems, new methods should be created and their confidence and accuracy checked.

One example of an innovative approach is using signs of life, signs of dwelling placement and signs of partnership for forecasting (calculation) demographic indicators. The indicators assessed by such methodology have the common name index. Suitability of the so-called index methodology has been checked when estimating the residency status of persons recorded in the Population Register of Estonia. The successful approach for checking residency encourages using a similar methodology for solving other demographic tasks as well. It is challenging to use the index methodology for solving the problems arising in organising a register-based census in a case when several register data might be of poor quality or the registration culture of the population is poor.

Methodology of building indexes for determining a specific part of population

In creating the index-based methodology, we use the approach known as the theory of fuzzy sets [11, 12].

The first step is to define the general population or research population that might have a different character. In the paper, three different types of general population were considered:

1. The set of points from persons' population **P** (coverage control and residency control);
2. The set of pairs from the same population **P**×**P** (partnership control);
3. The set of pairs from two different populations **P**×S (actual place of residence control).

In all cases, a subset of the general population (kernel) with special properties is defined (residents; confident pairs; actual places of residence). Belonging to a kernel can be characterised by an index: points j belonging to a kernel have the value of index $\boldsymbol{I}(\boldsymbol{j})$ equal to 1, other points of the general population the value 0.

The problem, common for all tasks, is to determine the points of research population belonging to the kernel. The task can be formalised as assessing the value of index $I(j)$. In different tasks, the index had the name of the residency index $R(j)$, the partnership index $P(h,j)$ and the dwelling placement index $D(j,s)$.

For solving the task, it is necessary to make some assumptions.

1. It is assumed that for a part of points their status (do they belong to the kernel or not) is known – this is the test population consisting of points from kernel **R** and points not belonging to kernel **N**.
2. There exists (can be defined) a set of binary variables – signs $L(j,i)$ correlated with the index $I(j)$. These variables can be defined in several ways, for instance:
   a. Activity in a register during a certain year;
   b. Having a certain status in a certain year (is married; is a legal inhabitant of a dwelling);
   c. Some information derived from the analysis of big data.

Using the two test populations **R** and **N**, the formula forecasting the status of a point by signs $L(I,j)$ has been derived (see also formulae (1), (5) and (6)):

$$I(j) = \sum_{i=1}^{m} a_i \, L(i,j). \tag{7}$$

The coefficients a_i for signs might be defined using traditional methods of multivariate analysis (logistic or linear regression or discrimination). The additional task is to define threshold c so that point j has been settled to the kernel of population if $I(j) \geq c$ and out of kernel in the opposite case.

Innovative methodology for defining model coefficients

There also exists another way for determining the values of coefficients a_i.

Let **R** be the part of the test population belonging to kernel and **N** the part of the test population not belonging to the kernel; let the sizes of these parts be $n(\mathbf{R})$ and $n(\mathbf{N})$. The expressions

$$\frac{\sum_{j\in \boldsymbol{R}} L(i,j)}{n(\boldsymbol{R})} \text{ and } \frac{\sum_{j\in \boldsymbol{N}} L(i,j)}{n(\boldsymbol{N})}$$

are equal to the average frequencies of sign $L(i,j)$ in sub-populations **R** and **N** of the test population. It is understandable that the higher the frequency of sign i is in set **R** and the lower it is in set **N**, the better predictor sign i is for index I. The ratio of these averages (see also (2)) is a coefficient characterising the impact of sign $L(i,j)$.

$$a_i = \frac{\frac{n(\boldsymbol{N})}{n(R)}\left(\sum_{j\in \boldsymbol{R}} L(i,j)\right)}{\left(\sum_{j\in \boldsymbol{N}} L(i,j)\right)} \tag{7}$$

Coefficients a_i are useable in the model for defining the value of indexes $I(j)$. The advantages of the coefficients a_i calculated in this way compared with the coefficients calculated using traditional methods of multivariate statistics are the following:

1. Such coefficients can also be calculated in the case when sign i has an impact for very few points only.
2. Using coefficients a_i there is no need to build different models for special groups of the general population (e.g. sex-age groups in the first example).

Some additional steps can be taken to improve the properties of coefficients.

1. Using instead of coefficients a_i the logarithms $q_i = \ln(a_i)$ also allows to use "negative" signs with the ratio (7) less than one.
2. Standardising the coefficients by the ratio of averages (3) changes the scale of indices closer to that of simple sums.
3. Truncating the calculated index so that it fulfils the condition $0 \leq I(j) \leq 1$.

Sequential indexes

In demographic calculations, it is common to calculate the indexes sequentially for sequential time-periods (years). Here, the stability of results is important. In this case, it is suitable to use for calculating the index for year k a linear combination containing also a term that characterises the status of point j last year, see [4]:

$$I(j,k) = dI(j,k-1) + g\sum_{i=1}^{m} a_i\, L(i,j,k)$$

where k indicates the year.

REFERENCES

1. Tiit E.-M. (2014). 2011. aasta rahva ja eluruumide loendus. Metoodika. Tallinn: Statistikaamet.
2. Tiit E.-M. (2012). Assessment of under-coverage in the 2011 Population and Housing Census. Quarterly Bulletin of Statistics Estonia, 4, 12, 110–119.
3. Tiit E.-M., Meres K., Vähi M. (2012). Assessment of the target population of the census. Quarterly Bulletin of Statistics Estonia, 3, 79–108.
4. Tiit E.-M. (2015). The register-based population and housing census: methodology and developments thereof. Quarterly Bulletin of Statistics Estonia, 3, 15, 42–64
5. Maasing E. (2015a). Eesti alaliste elanike määratlemine registripõhises loenduses. http://hdl.handle.net/10062/47557
6. Maasing E. (2015b). First results in determining permanent residency status in register-based census. https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=/149296295/170626623/Maasing_Abstract.pdf.
7. Zhang L.-C., Dunne J. (2015). Census-like population size estimation based on administrative data. https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=/149296295/170627702/Zhang_Abstract.pdf.
8. Tiit E.-M. (2015). Residence testing using registers – conceptual and methodological problems. https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=/149296295/170626640/Tiit_Abstract.pdf.
9. Tiit E.-M., Maasing E., Vähi M. (2017). Residency index – a tool for measuring the population size. Acta et communicationes Universitatis Tartuensis de Mathematica (accepted).
10. Rahvaarv ja rahvastiku koosseis. http://pub.stat.ee/px-web.2001/Database/Rahvastik/01Rahvastikunaitajad_ja_koosseis/04Rahvaarv_ja_rahvastiku_koosseis/04Rahvaarv_ja_rahvastiku_koosseis.asp.
11. Zadeh L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353.
12. Klaua D. (1965). Über einen Ansatz zur mehrwertigen Mengenlehre. Monatsb. Deutsch. Akad. Wiss. Berlin, 7, 859–876

Address for correspondence:
Ene-Margit Tiit
Institute of Mathematical Statistics
Faculty of Mathematics and Computer Science
University of Tartu, Tartu, Estonia
J. Liivi 2–513, 50409 Tartu, Estonia
E-mail: ene.tiit@ut.ee