ASSESSMENT OF THE TARGET POPULATION OF THE CENSUS

Ene-Margit Tiit
University of Tartu, Statistics Estonia

Koit Meres ***Mare Vähi***
Statistics Estonia ***University of Tartu***

What is the target population in statistics? What is the sample? What is the target population of a census? How accurate is the population figure established by a census and which statistical models can be used to correct it? The article provides a methodological overview of the determination of target population and the correction of under-coverage, and outlines the options for correcting the population figure.

The concept of 'target population'

'Target population' is a statistical concept denoting the whole set of all objects surveyed. While only a part of the target population is studied in case of sample surveys, the objective of the surveys is always to gain information on the target population. This is achieved through a suitably defined (designed) sample and rules of generalisation adequate for this design. Sample surveys are very common nowadays and provide a significant portion of knowledge about society and economy. However, censuses also play an important role alongside sample surveys. In case of censuses, all objects of the target population are surveyed, which means that the sample and the target population overlap (they are the same). Population census is the best-known and most important type of census, where the goal is to collect direct and immediate information about all the residents of a particular country (region).

The target population of a census

The target population of a census is the entire population of the respective country or region. The population is continuously changing: people are born and die (for example, there are, on average, around 40 births and deaths every day in Estonia) and they also move in or out of the country. Therefore, very specific definitions are required for an unambiguous understanding of the target population. The target population of a population census is determined as at the census moment (critical census moment). In Estonia, the critical moment of the last census (the 2011 Population and Housing Census or PHC 2011) was at midnight on 31 December 2011. Even though this was a population and housing census, this article will focus only on one aspect of the census, namely the population census, which means that 'target population' hereby refers to the target population of persons, leaving aside dwellings and households. The target population of persons includes all the residents of Estonia who were alive and living in the country at the census moment. People who were born after or who died before the census moment are not included in the target population.

While the criterion of time is relatively clear-cut, the criterion of residence is much more difficult to establish when defining the target population. In the past, the target population of a census has been approached in two different ways. One option is to enumerate the persons present in the country, i.e. the so-called factual population. This is done by identifying all the persons who are located on the territory of a given country at the census moment, including any short-term visitors in hotels, on trains and ships (in the territory of that country). The enumeration of the present population requires a relatively brief census period (one or two days, or up to 10 days in exceptional cases). The second option, which is preferred today, is to enumerate the permanent population. This group is more difficult to define, but with this option the census can be carried

out over a significantly longer period of time and will also provide more useful information for future application.

Permanent population

Permanent population includes all persons who reside in the given country at the census moment. People are considered as permanent residents of a country if they have lived in that country for at least one year (12 months) or if they, despite having lived there for a shorter period, intend to live in the country for at least 12 months. It is irrelevant whether a person is a legal or illegal resident in that country. However, there are international principles concerning certain special cases where people live or function in several countries (so-called 'transnational persons'). If a person has a family in one country, while he or she works in another country, but spends most of his or her free time with the family, that person is considered to be a permanent resident in the same country as his or her family. This definition also applies to persons who have been working in another country for more than one year. However, if a person is studying at a university or a post-secondary vocational school and the study period is at least one year, he or she is counted as a permanent resident at the location (country) of studies, irrespective of the frequency of his or her visits to relatives (parents) in another country. Separate rules have been established for diplomats, the staff of foreign missions, and participants in military missions; they are generally treated as permanent residents of their home country.

The main challenge in defining the permanent population is the fact that it is difficult to ascertain whether a person, who left a country less than a year ago, intends to remain abroad for at least 12 months or not. This is easier to do in the case of immigrants, because they (or their household members) can be asked this question during the census. The departed, on the other hand, are generally impossible to contact. Even though many of them have relatives in their homeland, the latter may not be aware of the long-term plans of the departed persons and the information collected from the household members is not always complete or correct.

Why the current preference for surveys of the permanent population? This is necessitated by the very high degree of mobility, which can cause relatively large variations in the factual population over a short period of time. The population of tourist areas can increase several times on a seasonal basis; the arrival of a large cruise ship in a small town can significantly boost its population; university towns become less densely populated during summer breaks, and so on. As the target population of a census is normally used as the basis for further population statistics (otherwise, such an expensive survey would not be economically justifiable), it should be as stable as possible. This is the reason why the permanent population is currently much better suited for population statistics than the factual population. In earlier times, when most people were settled in one place, the difference between the permanent and factual populations was marginal and had no significant impact on population statistics.

Nevertheless, the use of permanent population in population statistics also has its problems and questionable aspects, with the criterion of residence being the main source of confusion. Differently from a traditional census, respondents are asked to provide subjective information (whether they intend to stay in the country for a certain period of time) in addition to the usual objective information (the length of time they have stayed in the country). Any statements about intentions can be considered sufficiently reliable only if the answers are provided by the person in question. However, if the answers are given by another person (such as a household member, which would generally be acceptable in a census), such answers may not be valid. Unfortunately, it is inevitable that another person gives answers on behalf of a person who is (temporarily) absent from his or her place of permanent residence. As a result, it can be fairly difficult to differentiate between temporary absentees (away for less than 12 months) and those who have actually left the country (intending to remain abroad for over 12 months or on a permanent basis), based on the statements of respondents alone.

In the majority of previous censuses conducted in Estonia, both the permanent and factual population have been determined. In case of previous censuses, the statistics of vital events

were based on the factual population, but this has changed over time. In 2000, the majority of results were presented in relation to the permanent population. This was a time when the difference between the permanent and factual population was at the all-time high, with the permanent population exceeding the factual population by more than 13,000 people (nearly 1% of the population).

Censuses do not provide the accurate population figure

In principle, a population census should provide an objective picture of the population – including the exact size of the permanent population – that is not influenced by any external factors. However, this is only the case if all permanent residents of the country in question are enumerated, duplicate enumeration is avoided and enumeration of persons who should not be enumerated is prevented. Unfortunately, this is usually impossible.

One of the key indicators of accuracy is coverage, which shows the proportion of enumerated persons (L) to the number of persons subject to enumeration, i.e. the target population (N). Coverage is characterised by the ratio L/N, with a value under 1 indicating under-coverage and a value over 1 indicating over-coverage. Over-coverage can be significantly reduced in situations where the population is identifiable by personal identification codes. However, the main problem for censuses today is under-coverage, which shows the portion of the target population that was not actually enumerated. Under-coverage can be expressed by the ratio (N – L)/N, usually given in percentages.

Organisers of population censuses throughout the world agree that it is becoming increasingly difficult to reach people during a census. This is caused by a variety of reasons, but there are two that stand out – firstly, increased mobility of people, who often live and work in different places and even in different countries, with the accompanying increase in the variety of family and household types; and, secondly, a greater need for privacy, and unwillingness to disclose one's information to others (enumerators, public authorities). The fear that census data could be used against the respondent is still alive, despite information campaigns and security measures.

Accuracy of census results in previous censuses in Estonia

There are several ways to assess the accuracy of a census. If there have been no drastic population events between two censuses, the accuracy of a census can be assessed by comparing the continuously updated statistical records, which are based on the previous census, with the results of the new census. In essence, this is a measurement of consistency between the data of two consecutive censuses. This method helped to estimate the accuracy of the 1934 census as well as the accuracy of censuses conducted in the Soviet period (1970, 1979 and 1989). The results indicated a good accuracy in the 1934 census (under-coverage was about 1,500 persons, largely due to delays in the registration of newborns) and in the 1979 census (the difference was less than 1,000 persons).

Another method for assessing the accuracy of a census is a follow-up survey. In this case, a sample-based follow-up survey is used to identify persons who were not enumerated during the main census, and the resulting figure is extrapolated to the total target population in accordance with sampling rules. This method was used in the 2000 census. The results indicated under-coverage, i.e. a part of the target population was not enumerated. The estimated under-coverage was 1.2%, which was the minimum rate of under-coverage according to the author of the survey. Thus, it was clear that the target population established as a result of the census differed from the actual population size by at least 15,000 persons. However, this information was not used to adjust the indicators of population statistics.

The third option for assessing the accuracy of a census is to use additional information (such as registers). In principle, a comparison of the number of enumerated persons with the number of active entries in a representative register covering the whole population should allow an estimate of census coverage and, to an extent, of some other quality indicators. However, it is clear that

registers can only be used to assess the quality of a census if the registers themselves are of sufficiently reliable and high quality.

Options and methods for reducing under-coverage

There are several options for reducing under-coverage in a census, depending on the availability of additional information. However, these options have not been used very often in previous censuses. One option would be the use of weights (like in sample surveys). For example, if it is established that 3% of the residents of a settlement were not enumerated, each resident of that settlement is assigned a weight of 1.03 to correct the total number of residents in that settlement, while retaining exactly the same distribution by sex, age and other measured parameters as was established during the census. This method provides (more or less) accurate results in situations where enumeration gaps are completely random, i.e. not dependent on sex, age or other parameters of the residents – for instance, if the data from one enumerator are missing and the population characteristics in the area of that enumerator do not differ from the general picture of the settlement. In most cases, however, there are slight differences between the groups that have been enumerated and those that have not been enumerated – it is generally younger and more mobile people who tend to be left out. Therefore, use of this methodology is often not feasible. It is also not advisable in case of small settlements.

Further options for correcting the under-coverage of a census are offered by state registers. Several countries with a highly efficient and complete system of registers have stopped organising conventional censuses, replacing them with census-like inventories based on registers. The Nordic countries – Finland, Sweden and Denmark – were the first to choose this path. Even though some other countries have followed their example, the number of countries that organised register-based censuses was still below ten in the 2011 census round. However, registers can be used for correcting and supplementing census data. The idea behind such measures is very simple. Assuming that each resident of a country has been entered in a register (or registers) and is deleted from the register in case of death or emigration, such a register would be an excellent source for supplementing census data. It would help to verify the list of target population (and the number of persons included), but it should be remembered that no single register contains data to cover all census questions. Therefore, other sources have to be used to find answers to some of the questions in census questionnaires. The richest source of information is usually the population register, available in many countries, as well as various social security and health care registers. Many countries have registers today, but the quality, coverage and interoperability of those registers have rarely been analysed or assessed. There could also be problems with data protection – even though in Europe it is permitted (as an exception) to link personal data for statistical purposes, stricter regulations may apply in certain countries.

The dilemma facing census organisers

Organisers of modern censuses are faced with a dilemma – whether to use the number of enumerated persons as the official population size or to adjust this number, especially in case of considerable under-coverage. Previous surveys did generally not have this problem. Firstly, random over- and under-coverage were balancing each other out and, secondly, there were no alternative information sources (e.g. registers). It is likely that mobility was lower and people were possibly more law-abiding, which resulted in smaller errors, while quality standards were lower.

Problems are caused by the use of 'as is' census data as well as by adjustment of the data.

- *The main problem in case of using census data 'as is' is the known inaccuracy of population statistics. Under-coverage by 1–2% could result in significant shifts in the distribution of other parameters. For instance, an age-sex group could appear 5–10% smaller than in reality, or a region could appear to have significantly fewer inhabitants than it has in reality (assuming that under-coverage was at least partially caused by the*

actions of enumerators). Inaccurate population data lead to inaccuracies in important population indicators – fertility and mortality indicators and even economic indicators (GDP per capita).

- *Adjustment of census data also causes a myriad of problems. There is no internationally recommended standard methodology for data adjustment, which means that the methodology has to be developed separately in each country depending on available resources (data sources). Secondly, the methodology has to be sufficiently transparent and comprehensible to prevent any suspicions of political bias in the estimates. Thirdly, population figure is not the only value that needs adjustment. Based on alternative data sources, other important census parameters have to be determined for persons who have been added.*

Due to all these problems, adjustment of census data with the purpose of correcting population size has rarely been used in practice. However, one could assume that this will be done to a greater degree with the census data of 2011. The Latvian Central Statistical Bureau has already adjusted the census-based population figure on the basis of registers.

Under-coverage of the 2011 census and correction of population figures

There are clear signs of under-coverage in the results of the 2011 census of Estonia, published officially on 31 May 2012. This was confirmed by a number of messages received by Statistics Estonia as well as several media reports. In addition to the usual causes of non-enumeration (temporary absence from home, unwillingness to talk to an enumerator and disclose information, errors and omissions by enumerators), the 2011 census revealed a further cause for the non-enumeration of some people. The e-census, organised in the first stage of the census, was very successful, with nearly 66% of residents completing the census online. All these people also entered the details of their places of residence. Even though they were asked to enter the actual, not the registered, place of residence, some people still entered the registered place of residence (or another place) where they did not actually live. If the place of residence entered in this manner was actually used by a different household that did not participate in the e-census, there was a possibility that the household was not enumerated, because enumerators did not visit dwellings that had been properly enumerated during the e-census. There were many reports after the census about people who were omitted for this reason.

The census team now has a decision to make – either consider the number of enumerated persons as the official population size, even though it is known to be smaller than the actual population, or try to correct it. There are no standard guidelines from international organisations for such a situation. If the census team of a country decides to correct the census results, this decision will be accepted. At the same time, it is also possible to submit uncorrected population figures, despite the known under-coverage of the census.

As the registered methodology of the 2011 census in Estonia included the use of registers at different stages of the census, the correction of census results on the basis of registers would be legally acceptable. The issue of correcting census results was discussed on 25 June 2012 by the PHC Scientific Council and even though no final decision was adopted, the council members tended to support the correction of estimated figures. During the meeting as well as during the preceding and subsequent discussions, council members emphasised the need for caution in any decisions, a preference for as small corrections as possible (if the status of a person is in any way doubtful, he or she should not be included among permanent residents), the importance of the transparency and robustness of any decision methodology, and the need to give thorough explanations to the media.

The matter was decided at a meeting at Statistics Estonia on 29 August 2012. The census data will not be revised, but in December 2012 under-coverage data will be published for the whole country as well as by age group and local government unit. This will allow those interested to calculate the estimated non-offset (actual) population size in all these groups.

This kind of correction of population figures is unprecedented in the history of Estonian censuses. Although the follow-up survey of the 2000 census also revealed under-coverage, no corrections were made at the time. Indeed, this would have been impossible. Firstly, there was no acceptable methodology. Secondly, there were not alternative data sources, i.e. reliable and verified/audited registers. Also, the data protection regulations in force at the time did not permit the linking of the databases of different registers, as the encryption methodology required for this procedure was not yet in the application stage.

As a consequence, throughout the past 12 years, Estonia's reported population size has been slightly smaller than the actual population size. Assuming that the level of under-coverage of the census was 1.2–1.5%, it is likely that there were around 20,000 people more in Estonia at the start of the period (the beginning of the 2000s) than was reported on the website of Statistics Estonia (SE). Over the years, the difference between actual and official population figures has decreased and it is even possible that the situation is now reversed. This is mainly caused by another population process with the opposite effect – namely, external migration, which has a negative balance during the period observed and has remained partially unregistered. It turns out that correction of the 2011 census estimates would also necessitate an evaluation of the results of the 2000 census, and possibly also some adjustment of the population figures from the intermediate period.

The preference for corrected population figures is based on a very natural desire to have as accurate a picture of the population as possible, as this would provide the optimal foundation for policy decisions on both state and local levels. In addition, there is also the need to harmonise the three different population figures that have been used in Estonia – population size without migration (according to SE), population size with migration (according to SE) and the number of Estonian residents (according to the Population Register). The differences between all those three figures are above 20,000 persons, i.e. 1–2% of the population. The differences originate from the factors described above – under-coverage of the 2000 census, and the issue of including/excluding the impact of migration. It is also notable that the three sources report different population sizes for different age groups. For instance, in case of pre-school children, the number of residents according to the Population Register (which gives the largest figure overall) exceeds the population figure of SE without migration, but is below the population figure of SE with regard to children aged 7–12 (primary education age). The number of young people, aged 25–30, is also smaller in the Population Register database than in the SE database. This indicates that all currently used databases contain inaccuracies, especially considering that none of them reflect unregistered migration. It is clear that none of the existing databases provide a perfect reflection of the actual population in terms of age-sex distribution and geographical location, although some of the databases could be relatively close to the actual population size as at the census moment of PHC 2011.

Estonia's options for correcting the population figures of PHC 2011

The 2011 census was preceded by a period of hard work to analyse and improve the registers. Compared to other countries, the registers in Estonia have both strengths and weaknesses.

- *In Estonia, the main registers containing personal data use personal identification codes for identification, which enables linking.*
- *An address standard (ADS) developed in Estonia enables the description of all addresses of dwellings and other important locations (e.g. workplaces) according to a uniform system.*
- *Estonia has a functional Population Register which continuously records vital events (births, deaths, registered changes of residence).*
- *Estonia has an education information system (EHIS) which contains data on all students, teachers, education documents and certificates; a health insurance register (Health Insurance Fund) comprising many sub-registers; a register of the Estonian Tax and*

Customs Board containing data on taxable persons; a social security register with data on various kinds of benefits and pensions, as well as a number of other registers (see also "Enumerators' activity after the Census", Quarterly Bulletin of Statistics Estonia, No 2, 2012).

However, alongside these positive aspects, attention should also be paid to disadvantages and weaknesses.

- *All Estonian registers are relatively new – most of them were established in this century and, therefore, there is limited experience of use and combined analysis (and thus of error detection).*
- *Some registers are not sufficiently updated. For example, people of certain ages could still be listed in the health insurance register, even if they have left Estonia.*
- *The main weakness of the Population Register, the main register in Estonia, is the difference between registered and actual places of residence. In up to fifth of the cases, people do not live at their registered address. This phenomenon is the result of a number of developments which started at the beginning of the 1990s when the Parliament abolished mandatory address registration as a relic of the Soviet period. Even though residence registration has again been made mandatory, many people are still unaware of this, believing it to be voluntary. Inaccurate residence registration is (in addition to simple laziness) facilitated by various local benefits and concessions (possibility to choose schools and nursery schools, pension supplements, travel fare concessions). All local governments, including Tallinn, are interested in having as many registered residents as possible. All of this has led to the situation where the actual geographic distribution of permanent residents in the country could differ significantly from the distribution according to the Population Register.*

Disregard for the requirement of residence registration also causes errors in estimates of actual population size (target population). People who do not consider it important to register their place of residence often also fail to register the fact that they are leaving the country, which means that, formally, they remain residents, even though they may have left several years ago. Such behaviour could also be caused by rational (self-interested) reasons – by formally keeping a place of residence in Estonia, people retain the right to receive certain services from the state. On the other hand, this behaviour can be interpreted as a desire to maintain ties to Estonia, with the prospect of eventually returning to the homeland.

Different methods for correcting census under-coverage

Clearly, enumerators cannot simply 'invent' the missing persons. The target population can only be supplemented with persons who have been entered into Estonian registers and who can be assumed to have been permanent residents of Estonia at the census moment.

A. The most natural method would be to analyse the persons who have been entered in the Population Register as Estonian residents but whose data are missing from the collected census data (Figure 1, p. 85).

B. Another option would be to analyse, in addition to the above group, those persons with an Estonian personal identification code who have been entered in the Population Register but who do not live in Estonia according to the same register (they live abroad or the country of residence is unspecified).

C. In addition to the persons in the Population Register, we could also analyse persons who have an Estonian personal identification code and are listed in another Estonian register.

Supplementary information on all these persons can be obtained from all (remaining) state registers. For the purposes of supplementing the data of the 2011 census, it would be practical to use those activities that were entered in the registers during 2011. Clearly, not all registers are equally suitable for this purpose. Some registers require activity on the part of the person and

documented proof of residence in Estonia, while it is not an absolute requirement for others. Variation in the reliability of register data can also be associated with the age of the person in question. For example, all children automatically have health insurance, which is not the case for the working-age population – they have to work or study in order to secure health insurance. Slightly more information can be gained from registered doctor's appointments covered by Estonian health insurance – such visits are less likely (although not impossible) among people who have moved abroad. If the register indicates that a person is a full-time student at an Estonian educational institution, this is a relatively reliable indicator that the person resides in Estonia. Similarly, social benefits allocated by a local government are fairly reliable indications of residence in Estonia.

There are, in principle, two different possibilities for using registers – expert assessments and statistical models.

Using expert assessments to specify the size of target population

After a thorough analysis of the contents and structure of registers, it is possible to identify the most reliable registers and to clarify the links between them, whereby information is transferred directly between registers. In this way we can also identify the registers that are more likely to contain information on persons who actually are not permanent residents of Estonia. By preventing over-amplifications caused by links between registers, it is possible to make expert assessments to decide whether a person was a permanent resident of Estonia at the critical moment of the 2011 census. In principle, it is possible to define assessment rules for different groups of persons (see Figure 1, p. 85).

As a rough assessment, a person in group A could be considered as a permanent resident in 2011 if he or she is actively represented in at least two sufficiently reliable registers; in case of groups B and C, the person should be represented in at least three such registers. Error estimates of the expert assessments can be made empirically, by applying the assessment to persons whose residence status is known (those enumerated as permanent residents or as departed).

The advantage of the expert assessment method is that it is easily understandable and does not require any specialist knowledge of statistics. The main disadvantage of this methodology is subjectivity. It is relatively difficult to prove the reliability or independence of the data of a particular register by using only 'soft' approaches. It is also impossible to verify whether the established rule gives optimal results, i.e. causes the smallest estimation errors.

Using statistical models to specify the size of target population

There is also another option where the expert's subjectivity has no significant impact on the result. It is based on statistical determination of the optimal differentiating rule. Next, the method of discriminant analysis is described. In this method, an algorithm is compiled on the basis of 'training data' to help make the decision. After the census, two clearly defined groups of persons can be identified. One group includes 'permanent residents' (P) – people who were residents of Estonia as at 1 January 2012 according to the Population Register and have been enumerated as permanent residents, whereas they answered the questions themselves. The other group includes the departed (D) – those who have been enumerated as departed (based either on their own response or the response given by family members) and who did not live in Estonia as at 1 January 2012 according to the Population Register. These two training groups are then used to establish the optimal differentiating rule, which can be based on a linear or logistic model. Next, we describe the establishment of a linear model. Information on activity in 2011 according to registers is used as the arguments (descriptive parameters) of the model, as was the case in expert assessments. The list of potential arguments could include, in addition to existence of entries in a register, also various combined parameters and indices established on the basis of registers, for example, to take into account the frequency or date of activities in a respective

register or its sub-registers. What is important is that the selection of arguments for the model and the assignment of weights must be done automatically. The algorithm functions by first selecting the strongest parameter in differentiating permanent residents from the departed. Another parameter is added at the next step, creating the strongest pair of parameters for differentiating the groups. This process is continued until the addition of new parameters no longer significantly improves the model. As the level of representation in registers is strongly dependent on a person's age and partially also on sex, it would be practical to develop separate rules for individual age-sex groups. For this purpose, all persons are divided into age groups, also taking into account sex in case of the working-age population. Group boundaries are defined on the basis of actual frequency of occurrence in different registers. The optimal number of groups turned out to be nine; the provisional names of the groups are listed below:

1. *Children (aged 0–6);*
2. *Students (aged 7–19);*
3. *Young men (males, aged 20–29);*
4. *Young women (females, aged 20–29);*
5. *Middle-aged men (aged 30–39);*
6. *Middle-aged women (aged 30–39);*
7. *Older men (aged 40–59);*
8. *Older women (aged 40–59);*
9. *The elderly (aged 60 and older).*

The number of required parameters ranged from four to seven for different groups (Table 1, p. 87).

Three parameters were established on the basis of data from the Health Insurance Fund. HK1 is a binary (no/yes) parameter indicating presence in the Health Insurance Fund register in 2011. HK2 indicates the number of sub-registers that a person belonged to. HK3 contains the most reliable information on a person's insurance status. There are also two indices for education. EH1 is a binary (no/yes) index, while EH2 also characterises multiplicity (for example, a higher rating is given to a person who is simultaneously teaching and studying). MTA reflects the fact of receiving income from an Estonian enterprise. Sotst1 refers to receipt of family allowances (which could describe both the child and the parent) and Sotst2 refers to receipt of social benefits. STAR indicates receipt of support or allowance from a local government. The value of the Mntam index is determined by the presence of the person in the motor third-party liability insurance register. The list of potential arguments, considered as multipliers for the model, was actually much longer – it included pensions, parental benefits, incapacity and disability benefits and so on, but these parameters did not add any new information to the model in terms of differentiating between groups.

In some cases, the model in Table 1 (p. 87) includes several parameters established on the basis of the same register. Since they do not have different (plus/minus) signs, this is not a case of multicollinearity (which reduces the accuracy of a model and complicates interpretation), but a reflection of the fact that the impact of being listed in the respective register is not linear. It is notable that presence in the registers of the Health Insurance Fund is the parameter with the highest discriminating value for all age groups. Indices established on the basis of the education information system are important as well – even though this register does not cover many middle-aged and elderly people. Family benefit has the greatest discriminating value among social benefits and, as expected, presence in the register of taxpayers (MTA) is a significant discriminating factor for all working-age people. Despite some doubts (it is claimed that foreigners also often prefer to take the driving test and to insure their car in Estonia), active entries in the motor third-party liability insurance register in the given year add further information on the status of a person as permanent resident, even though it is the parameter with the least impact in most of the models. The fact that a number of supposedly important registers were not included among discriminating parameters can be explained by the statistical relations between parameters – for example, if a person has been entered in a pension register or parental benefit register, it means

that he or she is also entered in the Health Insurance Fund register, meaning that the former register does not add any new information.

The automatically selected parameters of the model form a predictive function, which can be pictured (for each age-sex group) as a straight line between two points (Figure 2, p. 88). These points are 'average departed person' and 'average permanent resident', depicted as circles in Figure 2. The value of the predictive function, i.e. a point on this line, is calculated for each person in the training group. In the Figure, these points are marked by small ellipses. In case of permanent residents the value of the predictive function (prediction) is closer to 'average permanent resident', while the predictions for the departed are generally closer to 'average departed person'.

In this way, predictions can be calculated not only for members of the training group, but also for others (using the parameter coefficients in Table 1), who can then be counted as departed or permanent residents depending on the location of their prediction on the line. The structure of the model indicates that only persons who appear in the registers included in the model (at least in the important ones) can be counted as permanent residents, while people who do not appear in those registers or only appear in marginal registers can be counted as departed.

Establishing a threshold

In addition to the discriminating function, we also need a threshold value to be used as the basis for decisions (Figure 2, p. 88).

After the predictive line corresponding to the optimal model has been established, we need to specify the threshold. There is some liberty in threshold specification, but it cannot be a purely subjective decision. When specifying the threshold, it should be kept in mind that errors are inevitable when it comes to statistical judgements. The likelihood of judgement errors depends on threshold selection. This indicates that it would be practical to base the selection of threshold on the likelihood of judgement errors.

There are two potential errors in the solution of this problem.

- *The first type of error is made when a person is counted as a permanent resident, even though he or she has actually left Estonia (is permanently residing abroad).*
- *The second type of error is made when an actual resident of Estonia is counted as departed (residing abroad).*

A judgement rule can be created in two ways. With the first approach, both errors are deemed equivalent and we specify a threshold that would result in equal, as small as possible probability for both types of errors. The advantage of this method is the maximum accuracy of the resulting population size estimate.

The other approach is based on a cautious estimate of population size. In this case, the probability of the first type of error is kept as low as possible, which inevitably increases the probability of the second type of error. This results in a general under-estimation of population size. For instance, we could decide that the probability of the first type of error (i.e. incorrectly counting someone as a permanent resident) may not exceed 0.05. The specification of maximum permissible error is very common in statistics. It means that if, for instance, 10,000 persons are being assessed, 500 of them would be erroneously counted as permanent residents based on the model (but they would remain unidentified). If at the same time the probability of the second type of error is, for instance, 0.09, the judgement rule would count 900 permanent residents as departed and the population size estimate would be inaccurate by 400 persons.

One could ask: can we establish a judgement rule that does not lead to any errors at all? Unfortunately, this is generally impossible in case of statistical judgements. This is due to the fact that the values of all parameters are inevitably random. For instance, we can do nothing about the situation where there are no active entries in 2011 in any of the registers for some permanent residents, who are registered in the Population Register as Estonian residents and have been

enumerated (by an enumerator or by completing the census questionnaire in Estonia). Consequently, threshold 2 in Figure 2 (p. 88) is usually unachievable in practice.

Calculation of judgement errors

In principle, there are two ways to calculate judgement errors: use of theoretical distribution (usually normal distribution) to assess the position of objects; or empirical calculation of errors based on training groups.

In Table 2 (p. 90), we present threshold values for the linear model on the condition that the theoretical probabilities of the errors are equal. In addition to the theoretical probability of errors, we have also calculated the empirical probability of errors for such a threshold as well as the probabilities of erroneously counted persons in the training group.

An empirical error of type 1 occurs when a person belonging to the D (departed) part of the training group is assigned to group P (permanent resident) based on the judgement rule. In case of the used training data, the frequency of such an event is 11%, which is a relatively poor result. However, as the share of group D in training data is relatively low, this error has a limited impact on the training group, with only 0.14% of persons erroneously assigned to group P in this manner. An empirical error of type 2 occurs when a person belonging to group P (permanent resident) is assigned to group D (departed) based on the judgement rule. The relative frequency of this event is less than 4% and the share of persons erroneously counted as departed in this manner is 3.8% of the total training group. These calculations indicate that the empirical probabilities of errors differ from the theoretical probabilities by 2–3 times on average, whereas the probability of type 1 errors in particular is higher than expected. At the same time, it is clear that, due to the different sizes of groups D and P, the share of erroneously counted P-persons is not particularly large in the training data. As training data constitute a relatively large portion of the actual census data, these estimates also apply to a similar extent to actual census data.

Table 2 (p. 90) indicates that in case of all age-sex groups the threshold is between the average values of the two groups (as can also be seen in Figure 2, p. 88), being generally closer to the average of group P than the average of group D. The fact that a part of points in group P remained below the threshold (to the left in Figure 2) is caused by the abovementioned fact that not every enumerated person who lives in Estonia left a trace in the registers during the given year. It is almost impossible to reduce this error on the basis of registers. However, errors of type 1 (i.e. related to assigning people of group D into group P) can be reduced by moving the threshold.

Specifying a threshold according to the probability of a given error

Next, we look at the possibilities of reducing the probability of type 1 errors in judgements. This is directly possible in case of theoretical errors. In the following discussion, we assume that the probability of theoretical judgement errors in identifying permanent residents may not exceed 0.02 (Table 3, p. 91). This limit is called significance threshold. In order to find a judgement rule that satisfies this condition, we need to assign new thresholds to all age groups. The probability of type 1 judgement error increases with the new threshold in two age-sex groups (students and middle-aged women). However, it is not always possible to define a judgement rule (threshold) that corresponds to the given significance threshold. This is the case when groups are poorly differentiated, because group means are close to each other and objects are jumbled.

The new threshold (threshold 2) is in most cases located closer to the mid-point of group P than threshold 1. This reduces the probability of type 1 errors, but the probability of type 2 errors increases considerably. This does not apply to groups where the probability of type 1 errors was lower than 0.02 already with threshold 1 – in their case, application of the new threshold increases the probability of type 1 errors and reduces the probability of type 2 errors. The situation is the most complicated in the last age-sex group (the elderly) where groups are poorly differentiated (the difference between group means is only 0.1). In case of this group, it was not

possible to establish a threshold between the two averages so that the probability of type 1 errors would be 0.02. Instead, the significance threshold 0.1 was used (maximum permissible probability of type 1 errors). The probability of type 2 errors is almost 0.5 in this case as well.

In conclusion, it seems that, in case of a new judgement rule, the difference between empirical and theoretical errors is greater than with the first rule, but a good balance was achieved between the empirical probabilities of type 1 and type 2 errors. The share of persons erroneously assigned to group P is only one tenth of a percent in total training data, which is a good indicator overall.

Logistic models

Two logistic models were created. The first model assumes that the theoretical errors are more or less equal; the second model limited the probability of type 1 error to 0.02 (if possible). In case of the first logistic model, the number of parameters automatically selected by the program differed from the number of parameters chosen for the linear model (parameter HK was not used; additional registers (parental benefits, pensions, incapacity and disability benefits) were used for some single groups). In case of the second logistic model, the size of the training group was decreased a little, while retaining the general principles. The error estimates of these logistic models are outlined in Table 4 (pp. 92–93).

The logistic models decided to assign a somewhat larger number of potential persons to group P (permanent residents) (see Table 5, p. 93; the reference base (100%) is model Linear 1).

The most "critical" model (Linear 2) was also tested on the data on all actually enumerated permanent residents. The model decided that the share of non-residents is approximately 10%, which indicates that this model probably gives estimates with a slight offset. There was good match between the judgements of models Linear 1 and Logistic 2, where no significant offset occurred.

Verification of judgement results

The next step in a proper judgement process is verification of judgement results. In this case, this can be done by using register data from a period after the census moment, i.e. at the beginning of 2012. Clearly, not all permanent residents have had active registry entries made during this period, but one would expect that the differences in the representation rates of different groups will provide additional information on whether persons live in Estonia or not. We can use two aggregate parameters for this purpose: firstly, activity in 2012 based on data from several registers; and secondly, doctor's appointments and inpatient care between November 2011 and March 2012. Neither of these parameters was used in the definition of judgement functions.

Figure 3 (p. 94) indicates that nearly half (48%) of enumerated permanent residents of Estonia, who live in Estonia according to the Population Register (PR), have been active in registers. Register activity was almost half of that in the group of persons who were enumerated as permanent residents but did not live permanently in Estonia according to the Population Register. This is a very small group (about 0.5%) which includes, for example, exchange students. Its low activity level can be explained by barriers associated with the status of a non-resident. Persons who were enumerated as departed but live in Estonia according to the Population Register are about 10 times less likely to appear in registers than those enumerated as permanent residents, which is understandable. The lowest level of register activity in 2012 was observed in the group that was enumerated as departed and whose place of residence was not in Estonia according to the Population Register.

The final five columns reflect the persons who were not enumerated. The highest register activity, over 80% compared to enumerated permanent residents, was observed among those non-enumerated persons who were assigned to group P (permanent residents) on the basis of the stricter judgement rule (threshold 2). As expected, the register activity of those who were assigned to group P on the basis of the 'softer' judgement rule (threshold 1) was slightly lower

(90%). This somewhat lower level of register activity compared to enumerated persons is understandable – persons who were not enumerated are, by nature, less active than those who were enumerated. With both judgement rules, the register activity of people assigned to group D (departed) is quite similar to people listed as Estonian residents in the Population Register but enumerated as departed. The judgement rules were not applied to persons who did not live in Estonia according to the Population Register. Even though 1% of them have appeared in Estonian registers in 2012, this can be considered a random phenomenon and it would be impractical to start looking for permanent residents among this group (see also Figure 1, p. 85).

The second parameter, which is partially related to the previous one but still quite informative, is based on visits to the doctor (at Estonian medical care institutions, covered by Estonian health insurance) in the period of November 2011 to March 2012 (Figure 4, p. 95).

Figure 4 shows activity data for the same groups as Figure 3 and the general picture is quite similar. There is an even more marked similarity between the behaviour of people assigned to group P on the basis of the stricter judgement rule (threshold 2) and the behaviour of enumerated permanent residents; and a marked difference between the groups of the departed and permanent residents. The figure indicates that the groups defined on the basis of judgement rules exhibit clearer differences in terms of register activity in 2012 than the persons defined on the basis of census data.

The difference between the two judgement rules (threshold 1 and threshold 2), when applied to the actual group of non-enumerated people who were Estonian residents according to the Population Register, caused the re-assignment of about 1/6 of people in group P to group D. However, looking at total data, this constitutes less than 0.5% of the population.

Conclusion

It is possible to correct the estimate of Estonia's population size to a significant degree by adding to the enumeration results of PHC 2011 the persons who were not enumerated but who are very likely to be Estonian residents (census under-coverage), using data from Estonian registers, particularly the Population Register, in combination with reliable mathematical statistics methods.

Methodologically, it would be practical to:

- *Use judgement rules based on training groups (discriminant analysis) whereby the most suitable register parameters are programmatically selected from a wide range of parameters;*
- *Establish separate judgement rules for suitably selected age-sex groups;*
- *Use a judgement rule that restricts type 1 errors, in order to ensure the higher reliability of the result with regard to 'sensitive' conclusions;*
- *Apply the judgement rule only to non-enumerated persons who are Estonian residents according to the Population Register (as an exception, the rule is also applied to those whose census results are contradicting);*
- *Introduce the selected and applied judgement rule to the public (in printed and social media) with detailed comments;*
- *Disclose error estimates when presenting the results.*