STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 197–206, DOI 10.21307/stattrans-2021-011
Received – 02.07.2020; accepted – 14.09.2020

Estimation of the number of residents included in a population frame

Danutė Krapavickaitė[1]

ABSTRACT

A high migration rate of a population causes the number of a country's residents to become extremely volatile, which negatively affects the quality of population frames. The aim of the paper is to present a method for estimating the number of residents included in a study population frame. The method involves the cross-classification of a population register with other databases which contain information relating to the activities of the population elements in a given country. The estimates from an ongoing sample survey are applied to some of the cells.

Key words: population register, sign of life, sample survey, logistic regression model.

1. Introduction

High migration rates have become a prominent characteristic of everyday life. People travel to get to know new countries, other styles of life, for studies, work, economic and other reasons. Sometimes, when departing from a country, they do not know whether their departure is temporary or permanent. Moreover, migration may be declared and non-declared, the latter making the number of country residents unknown.

Lithuania has been witnessing a rather high migration rate with a shift in direction, when immigration started outweighing emigration in 2019 (Statistics Lithuania, Official Statistics Portal). Table 1 shows that the intensity of migration in Lithuania is critical as compared to the moderate net migration rates in the European Union as a whole.

Due to migration, the number of country residents becomes volatile. As a consequence, the quality of the population frames worsens. Of course, there are many other reasons reducing frame quality. The quality requirements for frames in social statistics are very important, they are presented on the Eurostat's home page

[1] Vilnius Gediminas Technical University, Lithuania. E-mail: danute.krapavickaite@vilniustech.lt. ORCID: https://orcid.org/0000-0002-2159-1167.

(Eurostat, 2019). The number of residents included in the frame of the study population needs to be estimated. If one data source is insufficient, the integration of information from several data sources is needed. The construction of the residency index (Tiit, Maasing, 2016; Lehto, Söstra, Tiit, 2018) is one of the methods used to estimate the number of the study population residents or country residents included in the frame. Such methods use data from multiple sources, which makes quality evaluation in multisource statistics closely interlinked with and as important as the estimation of the parameters themselves.

Table 1. Crude rates of net migration plus statistical adjustment per 1,000 persons

Year	Crude rates of emigration from Lithuania	Crude rates of immigration to Lithuania	Crude rates of net migration, Lithuania	Crude rates of net migration, EU-28
2019	10.5	14.3	3.8	3.2(*)
2018	11.5	10.3	-1.2	2.8
2017	16.9	7.2	-9.7	2.3
2016	17.5	7.0	-10.5	2.3
2015	15.3	7.6	-7.7	3.5
2014	12.5	8.3	-4.2	2.1
2013	13.1	7.4	-5.7	3.5

(*) provisional data

Note. The crude rate of net migration is equal to the difference between the crude rate of increase and the crude rate of natural increase (that is, net migration is considered as the part of population change not attributable to births and deaths). The value is expressed per 1,000 inhabitants (European Commission, EMN Glossary Search).

The aim of the current paper is to present a way of estimating the number of residents included in the study population frame using the cross-classification of the population register with some other available registers and applying sample survey data. The idea is to divide the main study population frame into a union of nonintersecting cells and to estimate the number of residents included in each of those cells separately. For this purpose, various statistical methods may be used: estimates from the ongoing sample surveys, statistical models, etc.

The total number of country residents should be no less than the number of country residents included in the frame because the population may also include residents not included in any of the registers used. The estimation of the residing population size is more complicated. It is studied in the paper by Heijden, Cruyff, (2020). A dual estimation system based on the cross-classification of two or more databases and the application of the log-linear model to the cell sizes is used in this paper.

This study is partially based on the results included in the Komuso project initiated by the European Commission, 2016-2019. The results of the project on the quality for multisource statistics are shortly described in Ascari et al., 2020.

2. Presentation of a method for the estimation of the number of residents included in the study population frame

In order to demonstrate the method, data of Statistics Lithuania are used for the case study. Specifically, the problem is to classify the study population frame into country residents and non-residents and to simultaneously estimate the number of frame errors. The study population frame is based on the domain of the Population Register, which contains data on the individuals having a personal identification (ID) code of the Republic of Lithuania (LT) and older than 16 on October 1, 2016.

The sign of life is an indicator demonstrating the activity of the element in the database. In order to find signs of life of its elements, the frame has been merged with the database of the Lithuanian Labour Exchange (LE) and the database of the State Social Insurance Fund Board (SI) of Lithuania as of October 1, 2016.

2.1. Description of the data

The study population frame is further classified by the following variables:

- The *identification type* of an individual: a valid ID document, an invalid Lithuanian ID document or no document at all.
- The *address type*: a person declared his/her official address at the municipality of Lithuania; a person declared his/her home address in Lithuania; a person declared a foreign address; and other cases. "Other" means frame errors: a house at the address declared was demolished, is a non-residential building, or other illogical cases.
- *Signs of life*. The construction of the third classification variable is based on the cross-classified groups of individuals belonging to the LE and SI databases with the following levels: SI only; LE only; neither SI nor LE; SI and LE.
- Based on these three variables, the study population frame is divided into 32 non-intersecting groups, as shown in Table 2. The study goes further – to identify the elements of each cell: whether they live in Lithuania and are country residents or whether they are non-residents.

It is assumed that the fact of belonging to the LE or SI database means that an individual actually lives in Lithuania. There is possibility of a person living in one country while working in another; however, such individuals are excluded from the

current study due to a relatively insignificant number of such cases and for the sake of simplicity.

Table 2. Classification of the population frame by three variables

Document type		Address type			
	LE/SI	Municipality	Declared	Foreign	Other
Valid	SI only	A1	B1	C1	D1
Valid	LE only	A2	B2	C2	D2
Valid	Neither SI nor LE	A3	B3	C3	D3
Valid	SI and LE	A4	B4	C4	D4
Invalid	SI only	U1	V1	Z1	Y1
Invalid	LE only	U2	V2	Z2	Y2
Invalid	Neither SI nor LE	U3	V3	Z3	Y3
Invalid	SI and LE	U4	V4	Z4	Y4

2.2. Estimation of the unknown cell sizes

According to our assumption about the signs of life, the elements belonging to cells A1, B1, U1 and V1 should live in Lithuania because they have a sign of life there. The elements of cells A2, B2, A4, B4, U2 and V2 should also live in Lithuania. Cells D1,…, D4 and Y1,…, Y4 mean obscure addresses and indicate frame errors. For individuals from cells C1, C2, C4, Z1, Z2 and Z4 their signs of life are in Lithuania, however, according to the frame, their addresses are foreign. No signs of life in Lithuania indicates that individuals from the cell Z3 live abroad; individuals from the cell C3 may also live abroad. Persons from cells U3 and V3 are assumed to live abroad. Cells A3 and B3 consist of the individuals who are inactive in both LE and SI databases; they may live in Lithuania or abroad. It is also possible that some of them are already dead. The number of the three latter kinds of individuals has to be estimated. Let us denote by I2 the subset of individuals from the group B3 who live abroad, and by M2 the subset of individuals from the group B3 who passed away; let I1 and M1 be the corresponding subsets of individuals from the group A3.

Let us denote the number of individuals belonging to the cells of Table 2 by the corresponding lowercase letters: m_1, i_1, m_2, i_2, signifying the sizes of M1, I1, M2, I2; letters with hats mean their estimates: $\hat{m}_1, \hat{i}_1, \hat{m}_2, \hat{i}_2$.

Based on this investigation, the contents of the cells of Table 2 are resettled to the cells of Table 3.

Table 3. Grouped cells of Table 2

Population	According to the frame		
	Live in Lithuania	Foreign address	Errors
Live in Lithuania	A1+A2+B1+B2+U1+U2+V1+V2+A3+B3+ +A4+B4+U4+V4-I1-I2-M1-M2	C1+C2+C4+ +Z1+Z2+Z4	D1+D2+D3+D4 +Y1+Y2+Y4
Live abroad	I1+I2+U3+V3	C3+Z3	Y3
Deceased	M1+M2		

The union of cells B1, B2, B3 and B4 (Table 2) usually serves as a basis for the sampling frame in social surveys at Statistics Lithuania. Using information on the reasons for non-response from the Lithuanian Health Interview Survey, 2014, the number of individuals i_2 of I2$\subset$B1+B2+B3+B4 living abroad is estimated using sample design weights:

$$\hat{i}_2 = \hat{t}_y = \sum_{k\in(B1+B2+B3+B4)\cap s} w_k y_k \ ,$$

here *s* means a sample drawn from B1+B2+B3+B4, and w_k are the sample design weights. The values for a binary variable y are $y_k = 1$ if the *k*-th individual has not responded to the survey because he/she lives abroad, while $y_k = 0$ otherwise. Assuming that individuals from B1, B2 and B4 participate in the survey diligently, non-response is due to B3. It follows that the set of individuals living abroad I2$\subset$B3, and $y_k = 1$ for $k \in$ B3. The number m_2 of people in M2 who passed away from the cell B3 may be estimated based on the reasons of non-response using the same ongoing survey exactly in the same way.

For people belonging to A3 it is not possible to identify signs of life in Lithuania. Some additional databases are needed, to which no access was available in the current study. Assuming that the distribution of individuals living abroad in B3 coincides with the distribution of those in A3, it is possible to estimate the number of individuals included in the frame who live abroad and who passed away in A3 using the logistic regression model created in B3$\cap s$. Using data from the same ongoing survey and B3, the logistic regression model for the same variable *y* is constructed:

$$P\left(y_k = 1 \middle| x_k\right) = \frac{e^{a+bx_k}}{1+e^{a+bx_k}}, \quad k \in \mathrm{B3}\cap \mathrm{s},$$

and estimated. Using an assumption that individuals from B1, B2 and B4 live in Lithuania due to their signs of life, we restrict ourselves with $k \in \mathrm{B3}$. In general, *x* should be a vector of auxiliary variables, but the data study shows that there is only

one variable available in the sample data and in A3 correlated with *y*: *age*. Then, the probability of living abroad for units from A3 is estimated:

$$\hat{p}_k = \frac{e^{\hat{a}+\hat{b}x_k}}{1+e^{\hat{a}+\hat{b}x_k}}, \quad k \in \mathrm{A3}.$$

By summing the estimated probabilities over A3, the estimator for i_1 is obtained: $\hat{i}_1 = \sum_{k\in \mathrm{A3}} \hat{p}_k$. It estimates the number of individuals living abroad from A3. In a similar way, $\hat{m}_1$ (the number of individuals m_1 who passed away from A3) is estimated.

3. Numerical results

For this study, data of Statistics Lithuania for 2016 are used (European Commission, 2016–2019). The ID code is available in the frame and other databases. It allows us to merge the lists, fill in Table 2 and get the following Table 4.

Using sample data from the Lithuanian Health Interview Survey, 2014 (7000 individuals, of whom 25 passed away, and 71 went abroad), the estimates of coefficients for the logistic regression model constructed for individuals to live abroad are $\hat{a} = -2.9743$ (s.e.=0.2893), $\hat{b} = -0.0387$ (s.e.=0.0074). The estimated coefficients of the logistic regression model constructed for individuals to pass away are $\hat{a} = -9.9250$ (s.e.=0.9663), $\hat{b} = 0.0706$ (s.e.=0.0132). An assumption is made that the model did not change from 2014 to 2016, and these coefficients are applied to the data for 2016. In order to estimate the quality of the logistic regression model, the known number of individuals in the sample who went abroad t_{abroad} is compared to its estimate $\hat{t}_{abroad} = \sum_{k\in \mathrm{B3}\cap s} \hat{p}_k$, obtained using a logistic regression model, by the relative absolute ratio: $r_{abroad} = \left| t_{abroad} - \hat{t}_{abroad} \right| / t_{abroad}$.

The sample data are randomly divided into two parts: logistic regression coefficients are estimated from one of the parts and applied to estimate the number of individuals living abroad included in the second part. The relative absolute ratio is calculated in the second sample part. The procedure is repeated independently 200 times. The average for the relative absolute ratio obtained for the estimate of the number of individuals who went abroad is equal to $\bar{r}_{abroad} = 0.199$. The average relative absolute ratio $\bar{r}_{away} = 0.390$ for the estimate of the number of individuals who passed away is obtained in the same way. Estimates from the sample survey $\hat{i}_2$ and $\hat{m}_2$ have estimated coefficients of variation $c\hat{v}\left(\hat{i}_2\right) = 0.12$, $c\hat{v}\left(\hat{m}_2\right) = 0.2$. The accuracy of

the estimates is not high because the proportions of the sampled elements who went abroad and who passed away are low.

Table 4. Frame coverage and domain classification

Document type	LF/SI	Address type			
		Municipality	Declared	Foreign	Other
Valid	SI only	21 266	1 146 209	2 372	57
Valid	LF only	7 974	115 077	1 146	19
Valid	Neither SI nor LE	38 547	1267 490	257 544	133
Valid	SI &LE	1 265	22 323	33	1
Invalid	SI only	445	13 079	139	5
Invalid	LF only	189	1 568	10	0
Invalid	Neither SI nor LE	2 282	26 155	40 146	121
Invalid	SI &LE	17	313	1	0

Table 3 is presented in the same numerical way as Table 5. It provides the estimates for the number of residents included in the study population frame and the number of non-residents belonging to this frame. The number of individuals correctly classified in the frame as residents and non-residents and the number of mis-classified study population frame elements can be derived from Table 5.

Table 5. Basis for measuring the number of residents included in the study population frame, frame coverage and domain classification

Population domain	Frame domain		
	Lithuanian address	Foreign address	Frame errors
Living in Lithuania	2 600 857	3 701	215
Living abroad	54 252	297 690	121
Deceased	9 090		

Table 6 accumulates the main results of the study and provides the frame coverage. It is derived from Table 5.

Table 5 shows the cross-classification of the frame information on the country of residence and adjusted information on the residence place with the errors on the margins. Summing over its lines, both diagonals and margins, the contents of Table 6 are obtained.

Table 6. Number of residents included in the study population frame and frame coverage measures

Indicator	Value	Frame proportion
Number of country residents in the frame	2 604 558	0.8809
Number of country non-residents in the frame	351 942	0.1190
Correct domain classification	2 898 547	0.9803
Domain miss-classification	57 953	0.0196
Frame under-coverage	336	0.0001
Frame over-coverage	9 090	0.0031
In-scope frame size	2 956 500	0.9999
Overall population frame size	2 956 836	1.0000

One more aspect in the estimation of the number of residents included in the study population frame is the accuracy of the estimate. It depends on the quality of the frame and on all the statistical methods included in the estimation.

4. Discussion

In our investigation, *study population* refers as a set of individuals who are currently alive and show their activities (signs of life) in various databases. Based on these activities, one can decide if an element really exists at a certain moment and if his/her activities show that he/she should belong to a certain study population domain (with higher or lower probability).

Frame is a relatively frozen list of population elements, based on a certain register or several databases. The element is included in this list formally, and errors may occur.

We aim to merge what we know about the population elements from the sources different than the frame with what is available in the frame.

Access to additional databases, such records on health services, pensions and allowances, driving licence renewal, etc., could provide more information on signs of life for the elements of the frame; therefore cells C3, U3 and V3 could be classified or estimated more accurately. Unfortunately, although such data exist, they were not accessible for this study.

The number of signs of life used in this study is very small: only two. It is expected that if more signs of life were used, a more accurate estimate of the number of residents included in the study population frame could be obtained.

The problem studied in the paper by Tiit and Maasing (2016) is the construction of the residency index for the Estonian population. It is a complex index based on data

from 27 databases and the same index for the previous year with the values between 0 and 1 calculated for every element of the "extended total population" of the country. It is calculated for every year and used to estimate the population size. With a proper threshold, this index can be used to classify the "extended total population" into the groups of elements belonging/not belonging to the study population. It may also show the over-coverage of any frame domain. In Estonian case, the residency index allows estimating the size of the resident population and population register over-coverage due to persons who have left the country without declaration. The problem solved and the method used in the said paper by Tiit and Maasing are quite close to our solution, although more complex and using more signs of life.

Based on the idea of signs of life it seems that for any frame for which over-coverage is expected, a certain classification method may be applied based on the data showing activities of the population elements. If it is successful, the estimation of the frame over-coverage will not cause any problems. Any methods for the estimation of the active population frame size can be used for calibration in sample surveys, register-based statistics and administrative data-based population censuses.

Acknowledgments

The author is thankful to the anonymous reviewers for their comments and suggestions regarding improvements of the current paper.

References

ASCARI, G., BLIX, K., BRANCATO, G., BURG, T., McCOURT, A., VAN DELDEN, A., KRAPAVICKAITĖ, D., PLOUG, N., SCHOLTUS, S., STOLZE, P., DE WAAL, T., ZHANG, L.-C., (2020). Quality of Multisource Statistics – the KOMUSO Project. *The Survey Statistician*, January No 81, pp. 35–49, http://isi-iass.org/home/services/the-survey-statistician/ (accessed January 2020).

EUROPEAN COMMISSION, (2016-2019). *ESSnet on Quality of Multisource Statistics – Komuso*, https://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics-komuso_en/ (accessed January 2020).

EUROPEAN COMMISSION, (2020). *EMN Glossary Search*, https://ec.europa.eu/home-affairs/what-we-do/networks/european_migration_network/glossary_search/ (accessed January 2020).

EUROSTAT, (2019). *Quality Guidelines for Frames in Social Statistics – QGFSS*, https://ec.europa.eu/eurostat/cros/system/files/qgfss-v1.51.pdf/ (accessed August 2020).

LEHTO, K., SŐSTRA, K., TIIT, E.-M., (2018). Index-based Methodology in Demographic Statistics. *The Survey Statistician*, January No 77, p. 45. https://isi-iass.org/home/services/the-survey-statistician/ (accessed January 2020).

STATISTICS LITHUANIA, (2020). Official Statistics Portal, https://www.stat.gov.lt/home/ (accessed August 2020).

TIIT, E.-M., MAASING, E., (2016). Residency index and its applications in censuses and population statistics. *Eesti statistika kvartalikri.* (*Quarterly Bulletin of Statistics Estonia*). No 3, pp. 41–60, http://www.stat.ee/publication-2016_quarterly-bulletin-of-statistics-estonia-3-16 (accessed January 2020).

VAN DER HEIJDEN, P. G. M., CRUYFF, M., (2020). Wider applications for dual and multiple system estimation. *The Survey Statistician*, January No 81, pp. 16–20, http://isi-iass.org/home/services/the-survey-statistician/ (accessed January 2020).