# Population Estimation Beyond 2021

Mārtiņš Liberts[1]

[1]Central Statistical Bureau of Latvia, e-mail: martins.liberts@csb.gov.lv

**Abstract**

Population statistics in Latvia are produced using register/model based methodology since 2012. Precision evaluation of register/model based statistics is an ongoing process. The paper summarise the activities done for precision evaluation so far. The current register/model based methodology has worked so far and we plan to use the same methodology for Census 2021. However, a long-term aim is to develop an alternative methodology for population statistics.
*Keywords*: Population estimation, Population census, register/model based statistics

## 1 Introduction

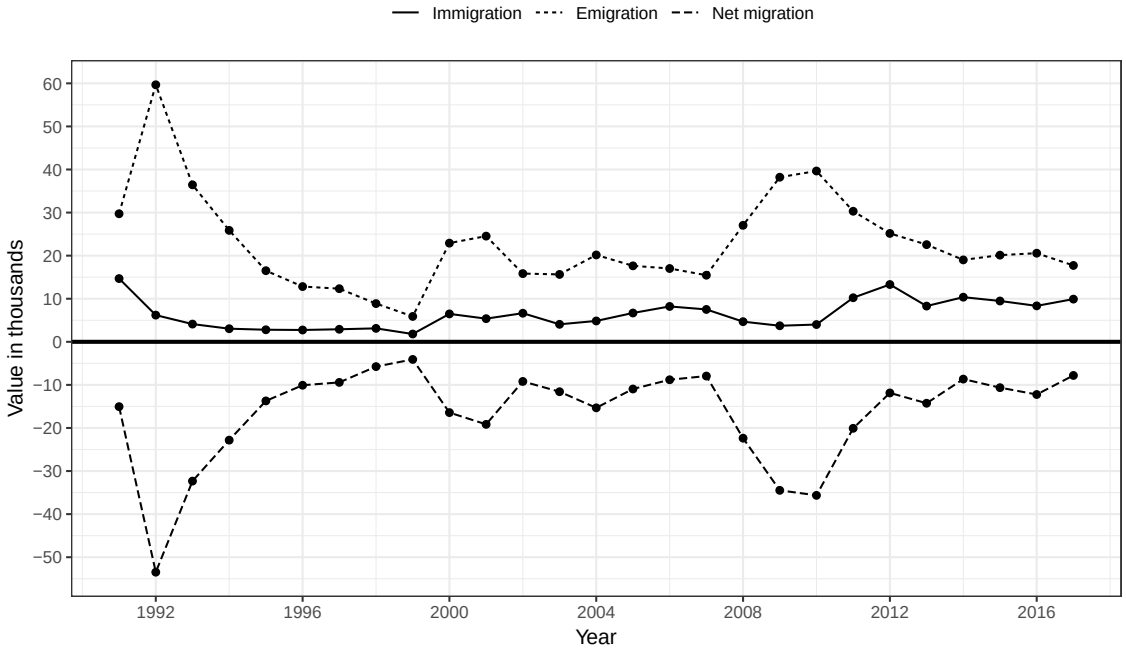
The population statistics of Latvia are produced using register/model based methodology since 2012 (Vaļkovska, 2012; Vaļkovska *et al.*, 2014; Aināre *et al.*, 2017). Even though there is Population Register in Latvia, it is not possible to produce population statistics using register based methodology only. The main reason being the significant over–coverage of the resident population in the register data.

The difference between the register population and the census population was 155 thousand (7 % from the census population) in 2011 (Aināre *et al.*, 2017, p. 1). The cause of this difference is twofold. Firstly there is still important emigration flow from Latvia ongoing with recent peak value in 2009–2010 when emigration reached almost 40 thousand emigrants per year (see Figure 1) which is 2 % from the total population. Secondly there is lack of incentives for emigrants to provide correct information to the Population Register about the place of residence. As the result for many emigrants the declared country of residence is still Latvia.

We can describe the problem as a statistical classification problem. The task is to classify all individuals from the register population into two classes – *de facto* residents and *de facto* non–residents of Latvia. We have solved the problem using a logistic regression model. The dependent variable in the model is a binary variable denoting *de facto* residents with `1` and *de facto* non–residents with `0` and independent variables are different binary variables describing individuals. The model is estimated using 2011 data, where dependent variable is taken from Census 2011 and independent variables are created from several administrative data sources corresponding as close as possible to the year 2011. Each following year the same set of independent variables is created and the model is applied to classify all individuals from the register population into residents and non–residents (Aināre *et al.*, 2017).

The estimation methodology including the model was developed during 2011–2013. The first estimates using the new methodology were published in mid 2013 (the results for 2012 were revised according to the new methodology). The original model with some adjustments (by adding or removing some dependent variables) has been used since then.

Figure 1: Long-term migration




*Source: Central Statistical Bureau of Latvia*

# 2 Precision evaluation

Precision evaluation of register/model based statistics is not straightforward. Usually some external data are necessary for the precision evaluation. Currently three approaches are used for the precision evaluation:

- Micro data of population statistics are linked with survey data or other administrative data not used for modelling. It is possible to detect persons which are residents of Latvia but which we have not included in the frame for population statistics. This approach is applied yearly. The range of errors is 1 % to 2 %. See Aināre *et al.* (2017) for more details. Unfortunately we can detect only one side of errors using this approach. We can not detect persons which we have included in the frame for population statistics but who are not resident of Latvia using this approach.
- Methodology to estimate model errors is under development now. The task is to estimate variability of model estimates (assuming the model coefficients are estimated). Bootstrap methodology is being applied for this task.
- We have organised two large scale sample survey – Micro census 2015 and International migration survey 2017–2018. The main aim of those surveys is precision evaluation of the population statistics.

## 2.1 Micro census 2015

Micro census was organised in 2015. The aim of Micro census was the precision evaluation of population statistics. Micro census was done as an independent sample survey. The target population of Micro census was all private dwellings of Latvia. The sample size was 14,996 dwellings. Two-stage cluster sampling design was used (Liberts, 2017).

Table 1: Estimates of population – total and by gender

| Gender | $\hat{\theta}_P$ | $\hat{\theta}_M$ | $\hat{\theta}_P - \hat{\theta}_M$ | $\hat{\text{me}}\left(\hat{\theta}_M\right)$ | p–value |
|---|---|---|---|---|---|
| Total | 1 949 510 | 1 912 299 | 37 211 | 28 837 | **0.011** |
| Males | 892 394 | 864 970 | 27 424 | 16 500 | **0.001** |
| Females | 1 057 116 | 1 047 328 | 9 788 | 19 211 | 0.318 |

The main aim of the data collection process was to list all residents of sampled dwelling. Gender and age of residents was recorded. This information allows to get estimates of resident population size in breakdown by gender, age, and regions.

The total over–coverage rate (weighted) was 21.1 %. It is quite high if compared with other sample surveys were dwelling sample is used. This was expected because the population frame was created including completely all private dwellings from different data sources. This was done to reduce under–coverage risk as much as possible. Most of over–coverage cases were because of unoccupied dwellings (72.8 %). The potentially unoccupied dwellings are excluded usually from the population frame for other surveys.

The total response rate (weighted) was 93.5 %. This is very high if compared to other usual surveys. It was possible to achieve so high response rate because of two reasons: questionnaire was very short and proxy interviews (with neighbours or local municipality) were allowed. So we can hope to have potentially low non–response bias.

The results of Micro census were compared with population statistics (excluding population of institutional dwellings). Since Micro census was carried out as a sample survey – the results of Micro census have sampling errors. It was taken into account when comparing the population statistics and the results of Micro census. Comparison was made with the help of hypothesis testing.

Analysing the total population, we can conclude, that the difference between the population statistics and Micro census of the population is 37 thousand (1.9 %), which is statistically significant, because the margin of error is 29 thousand (relative margin of error is 1.5 %). Micro census indicates that the total population is overestimated. The analysis of the results split by gender shows that the total number of men in the population statistics is overestimated. The estimates of number of women do not have statistically significant difference. See Table 1 where $\hat{\theta}_P$ is the estimate of a population parameter using the current methodology, $\hat{\theta}_M$ is the estimate of a population parameter using Micro census data, $\hat{\text{me}}\left(\hat{\theta}_M\right)$ is the estimate of margin of error for $\hat{\theta}_M$, and "p–value" is p–value from hypothesis testing (equality of $\hat{\theta}_P$ and $\theta$ is tested assuming $\hat{\theta}_M$ is an unbiased estimate of $\theta$).

Micro census results were rated as very valuable source of information for precision evaluation of population statistics. Some significant differences have been found between the current population statistics and the estimates from Micro census, however most of the differences are explainable and understandable. The results of this evaluation task show the direction of necessary improvements for the current methodology.

## 2.2 International migration survey 2017–2018

There was trial to estimate long-term international migration flows using Micro census. Unfortunately this trial was not successful. The main reason of failure was measurement errors. Micro census was done as one-wave survey. The field work of Micro census was organised during the 4th quarter of 2015. Respondents were asked to list residents of a sampled dwelling at three time points: 2015-01-01, 2015-09-01, and 2016-01-01. The listing of residents on 2015-09-01 was used for the population estimates as this listing was the closest

to the fieldwork period.

The listings of residents on 2015-01-01 (*past*) and 2016-01-01 (*future*) were used for migration estimation. Unfortunately the data collected about those time points were influenced by measurement errors. It was not possible to use Micro census data for reliable migration estimates.

Decision was made to organise International migration survey as a two-wave sample survey. The sample of 20,000 dwellings was drawn. The survey strategy is to monitor the sample dwellings in two time points. The task is to list the residents of sampled dwellings on two time points, namely 2017-12-01 and 2018-10-01. It will be possible to estimate international migration by comparing those lists (birth, death and internal migration should be excluded).

The data collection for the 1st wave has been done already. It was done from December 2017 till March 2018. The data processing is in process now. The data collection for the 2nd wave will be done at the last quarter of 2018.

# 3 Census 2021 and beyond

The current register/model based methodology for the estimation of population statistics has worked quite well. We have made precision evaluation of population statistics using Micro census in 2015. Precision evaluation of international migration statistics will be done using the results of International migration survey 2017–2018. The current plan is to use the same methodology also for Census 2021. So Census 2021 will be done as register/model based census in Latvia. However we have observed some drawback of the current methodology.

Firstly, the model used for population classification has been estimated using the data from Census 2011. So the question is – how long we can use this model? How to detect a time point when model fails to predict the current population? We do not have answers for those questions. But it is clear that it will not be possible to use the current model forever.

Secondly, the classification model works quite well for population size estimates. Unfortunately it fails to get good international migration estimates directly. The solution is to use external migration data and to estimate total emigration separately. Finally the results of model (probabilities) are adjusted to be in line with the external migration estimates. See Ainārе *et al.* (2017) for more details.

The long-term aim is to develop different methodology which would deal with those two drawbacks mentioned. There have been some attempts to achieve this aim.

We have tried to replace Census 2011 data with the data from a recent large scale sample survey data (for example, Labour Force Survey). This would allow to estimate the model using more recent data.

Another attempt was to replace supervise classification model (logistic regression) with unsupervised classification methods (for example, clustering). In this case it would be possible to discard Census 2011 data from the estimation phase.

Unfortunately none of those attempts have resulted with something reasonable. Work is in progress.

# 4 Conclusions

Population statistics in Latvia are produced using register/model based methodology since 2012. The same methodology will be used for Census 2021. The precision evaluation of register/model based statistics is not straightforward. Several approaches has been used for precision estimation. The current methodology has worked so far. However it is clear that

we will need to develop an alternative methodology. The main reason being that the current model is estimated using Census 2011 data. Census 2011 data becomes more and more outdated by time.

# References


Ainãre, I., Liberts, M., Zukula, B., Šulca, S., Vaļkovska, J., Opermanis, B., Jurševskis, A., Lece, K. & Ceriņa, A. (2017). Method used to produce population statistics. Methodological report, Central Statistical Bureau of Latvia, Riga, Latvia. https://www.csb.gov.lv/sites/default/files/data/EN/demstat_metodologija_eng.pdf.

Liberts, M. (2017). Methodology and results of the micro census in Latvia. In *Baltic-Nordic-Ukrainian workshop on survey statistics theory and methodology*. Statistics Lithuania, Vilnius, Lithuania, pp. 58–63. http://vilniusworkshop2017.vgtu.lt/wp-content/uploads/2017/07/Workshop-Proceedings-2017.pdf.

Vaļkovska, J. (2012). The number of Latvian residents estimation via logistic regression. In *Workshop of Baltic-Nordic-Ukrainian network on survey statistics*. Central Statistical Bureau of Latvia, University of Latvia, Riga, Latvia, pp. 198–202. http://home.lu.lv/~pm90015/workshop2012/papers/W2012_Poster_VALKOVSKA_JELENA.pdf.

Vaļkovska, J., Liberts, M. & Zukula, B. (2014). The estimation of population in Latvia. In *Workshop of BNU network on survey statistics*. Statistics Estonia, Tallinn, Estonia. https://www.stat.ee/public/yritused/BNU/Valkovska_Liberts_Zukula.pdf.