Residence testing using registers – conceptual and methodological problems

Ene-Margit Tiit

Helsinki, 28.08.2015

eesti statistika

Content

- What does mean residency testing?
- Estimating the under-coverage of census 2011
- How to go on?
- Residency model (Ethel Maasing)
- Stability of solution
- Residency index

WHAT DOES MEAN RESIDENCY TESTING?

Residence – why to test it?

Number of residents or population size is important for all countries, but also cities, towns and municipalities.

For long time the only way to get information about the number of residents was census.

Seemingly, from the time when different registers are in use, the situation has become easier – you can count the number of residents from registers, and that can be done principally at any time without interviewing the people.

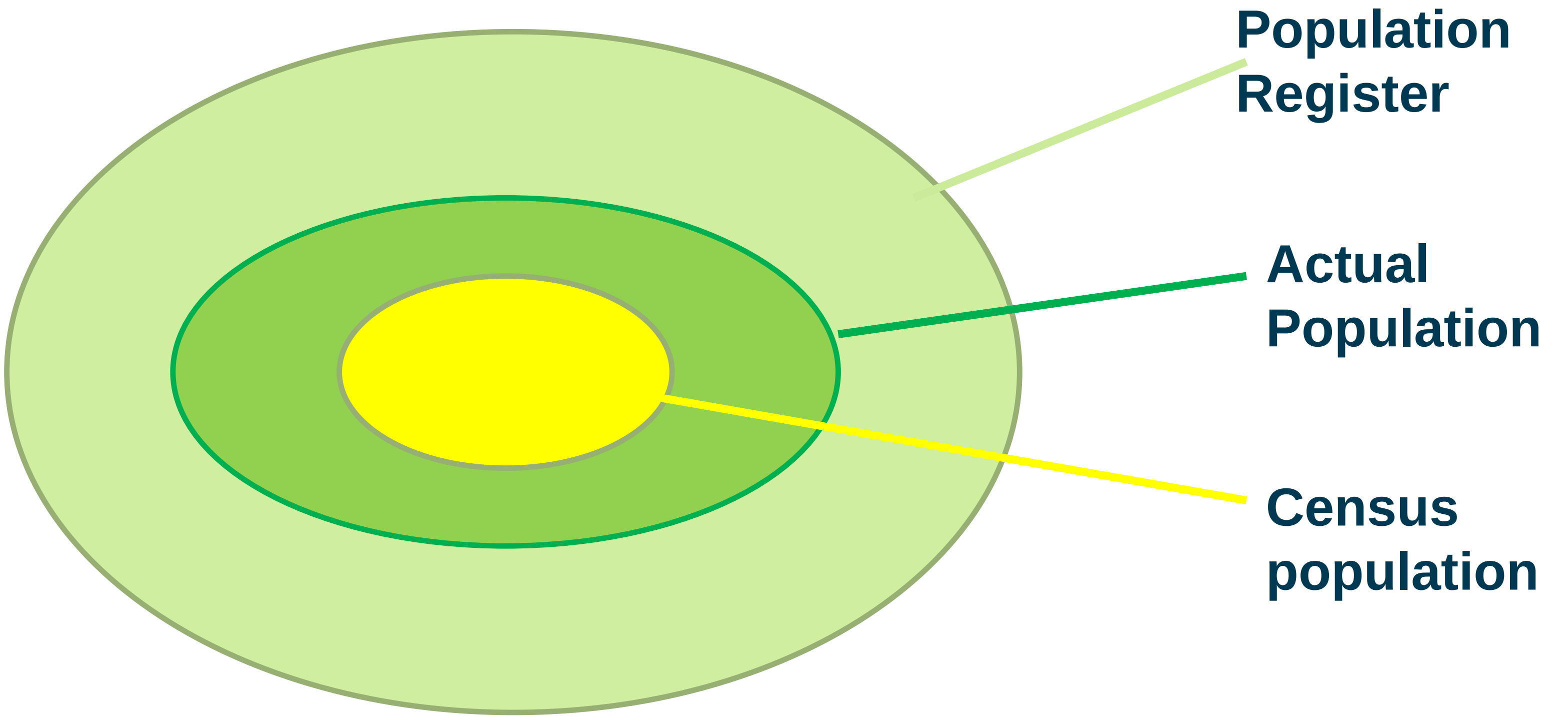
Several population sizes

But more sources of information sometimes complicates the situation.

For instance, in Estonia after census 2011 we had three different population sizes:

- Size of census population – 1 294 455;
- Population size calculated using registered population events and population size in 2000 – 1320 000
- Population size from Population Register 1365 000.

In some age-groups the difference between different estimates was almost 10%

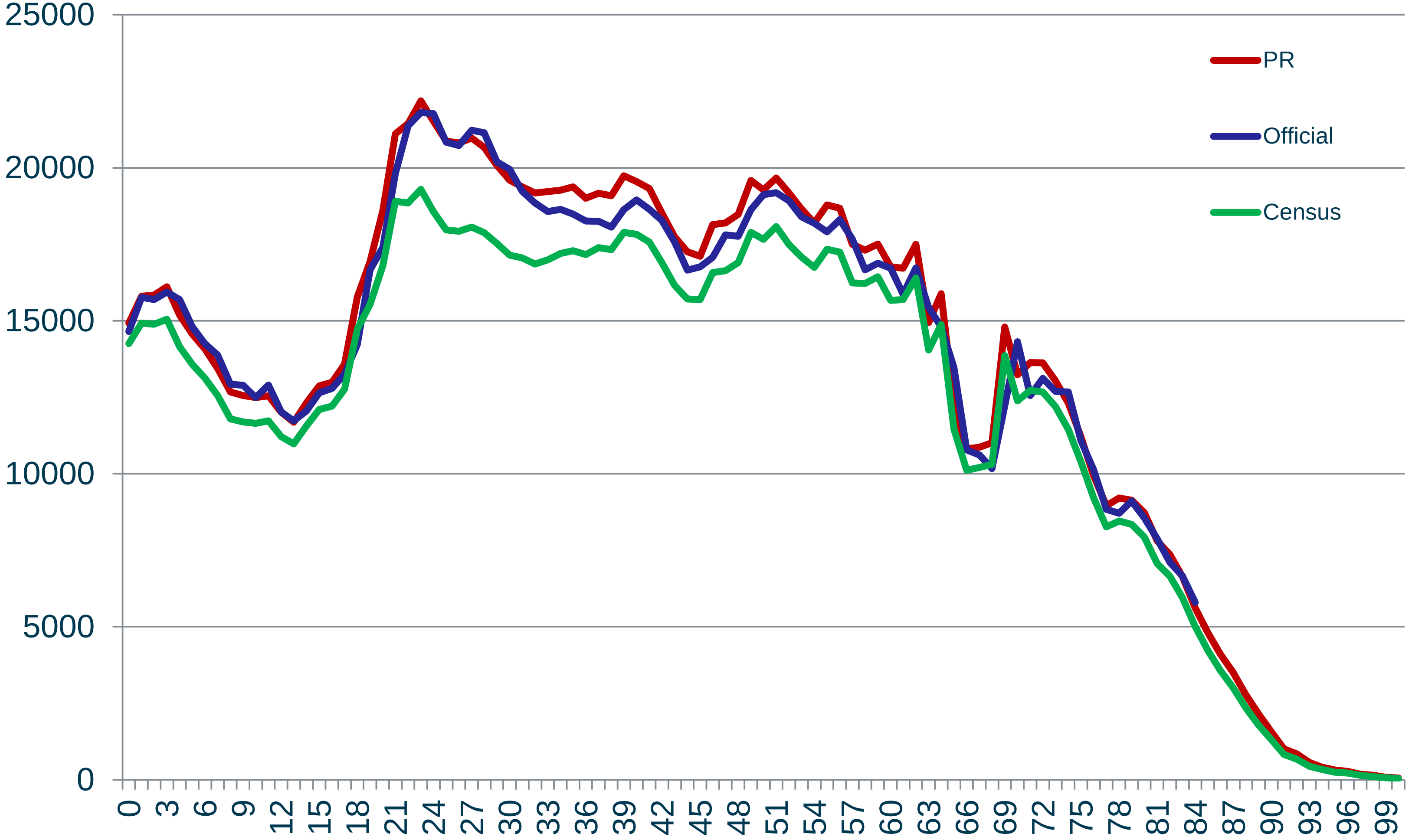


Population Register

Actual Population

Census population

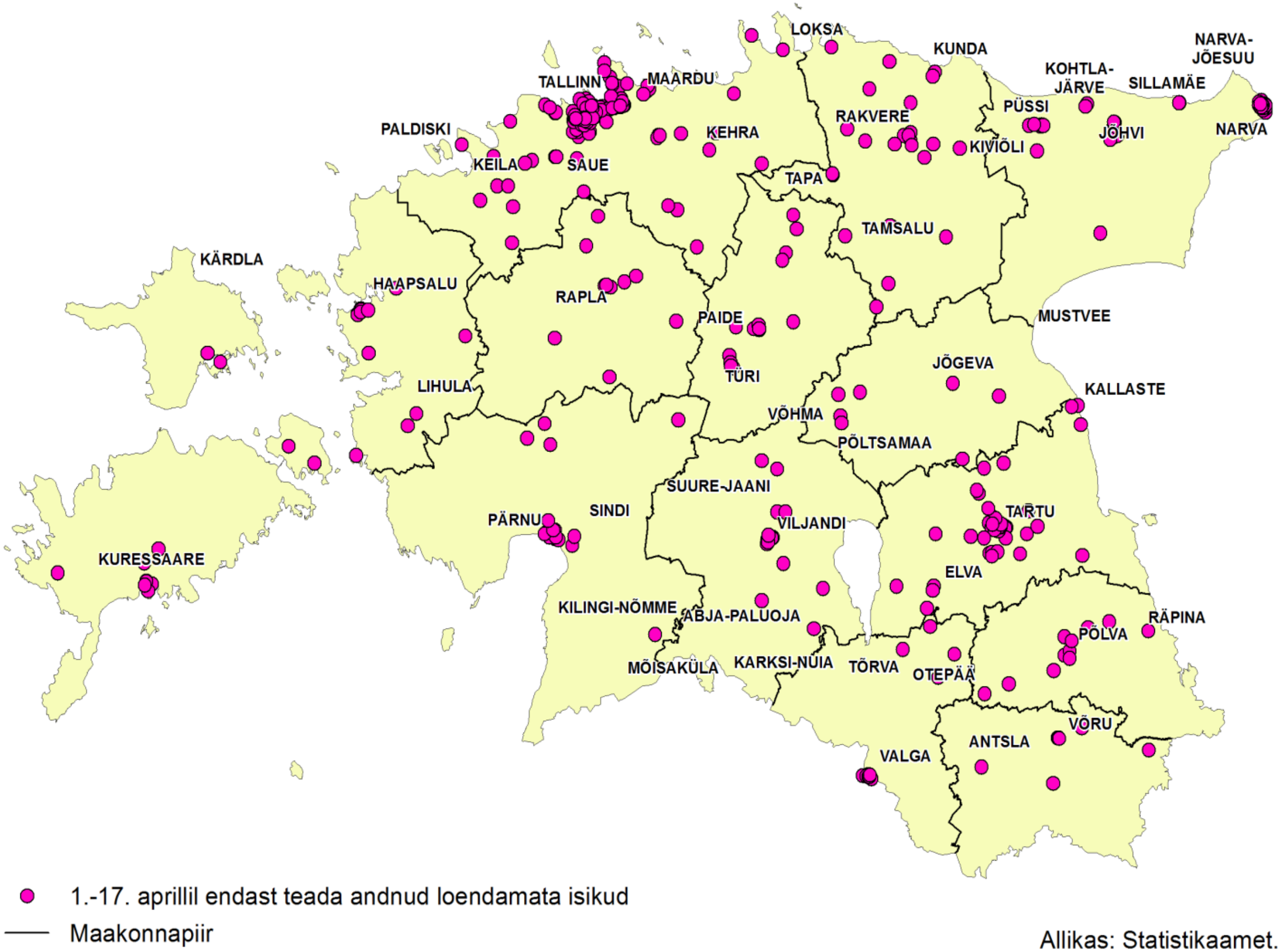
Age structure of Estonian population in 2011



eesti statistika

THE FIRST ATTEMPT TO TEST THE RESIDENCY IN ESTONIA: ESTIMATING THE UNDERCOVERAGE OF CENSUS 2011

Messages from people who were not enumerated in 2011

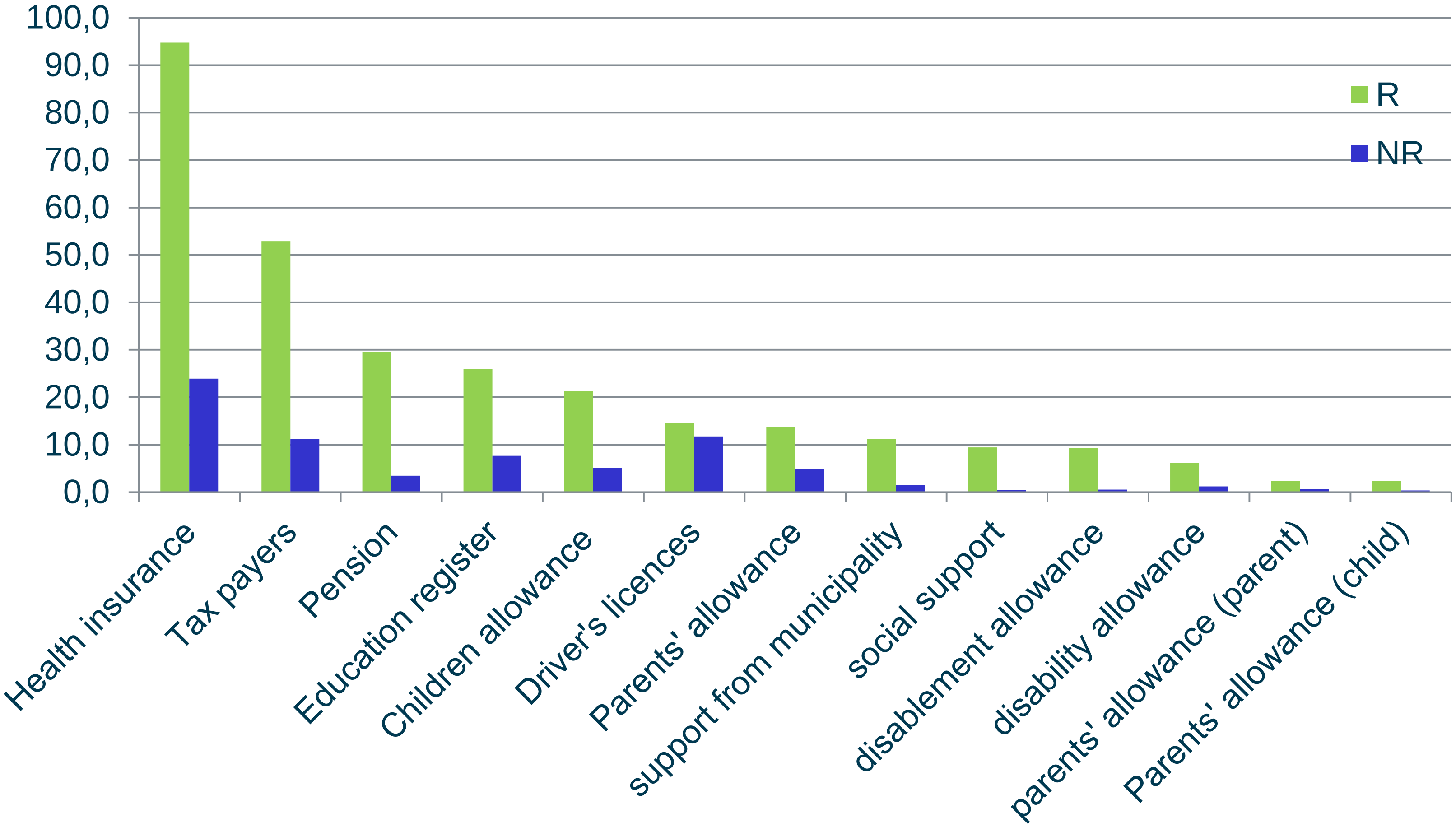


Allikas: Statistikaamet.

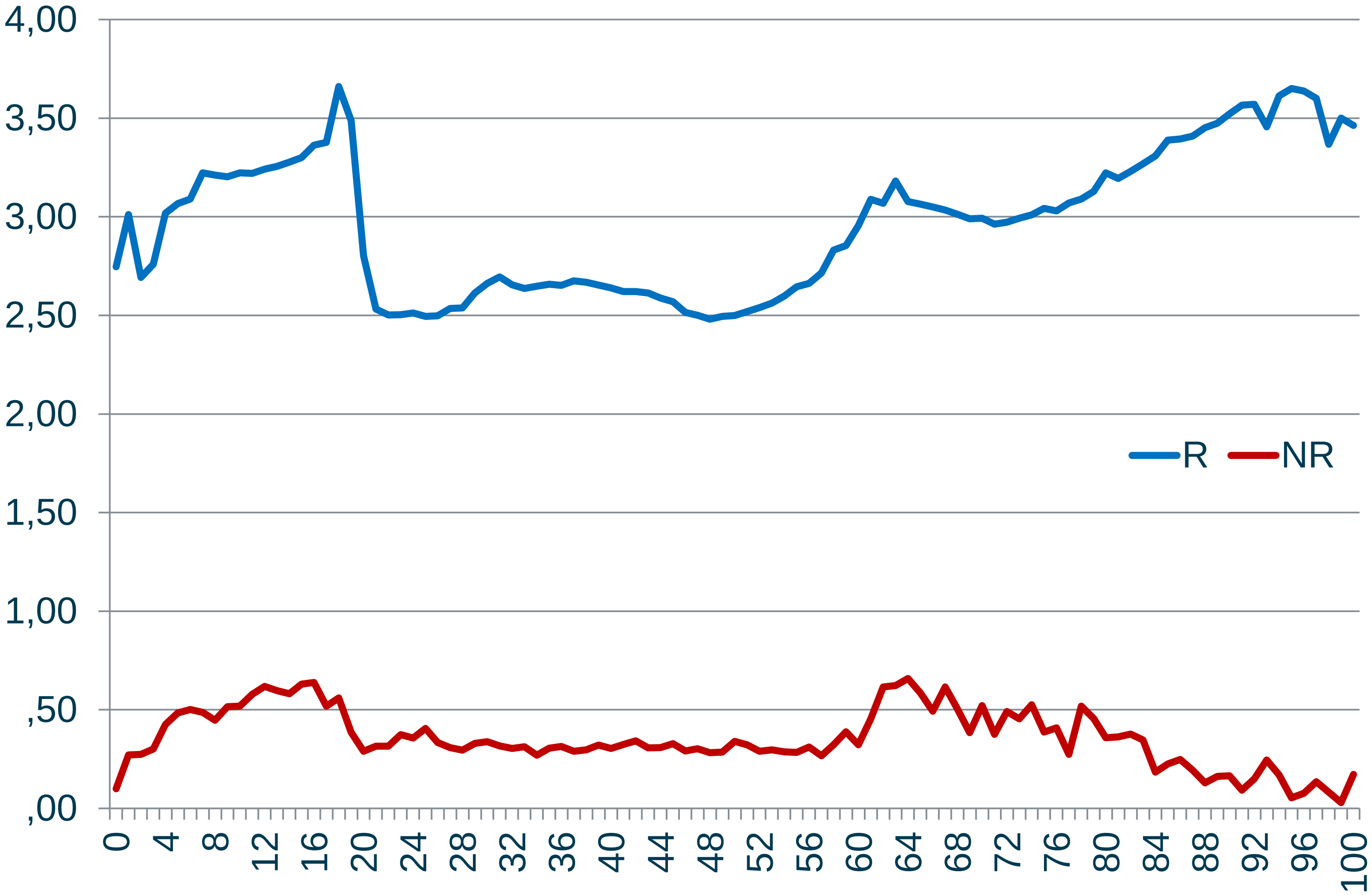
Estimating under-coverage of census

- Under-coverage was estimated in 2012 statistically, using logistical and linear regression, administrative registers as explanatory variables and census population as test-group [1—3].
- About 20 models for different age-sex groups were created and integrated.
- About 30 000 persons (2,3% of population) were added to census population to get the population for demographic calculations. Each added person was identified by his/her ID-code.
- Estimated errors (inclusion and exclusion) were less than 5%.

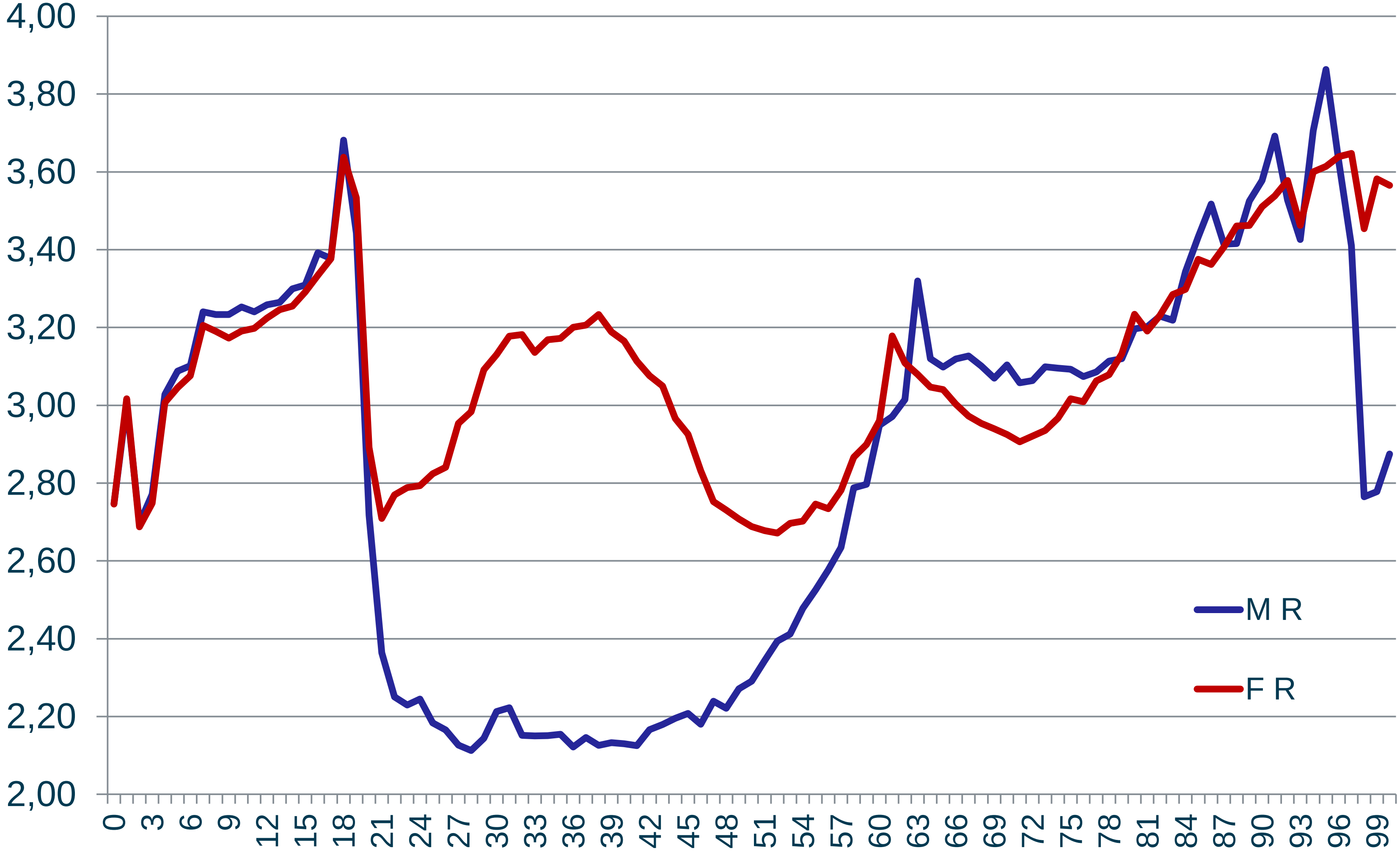
Activity in different registers, 2011 (% from the population group)



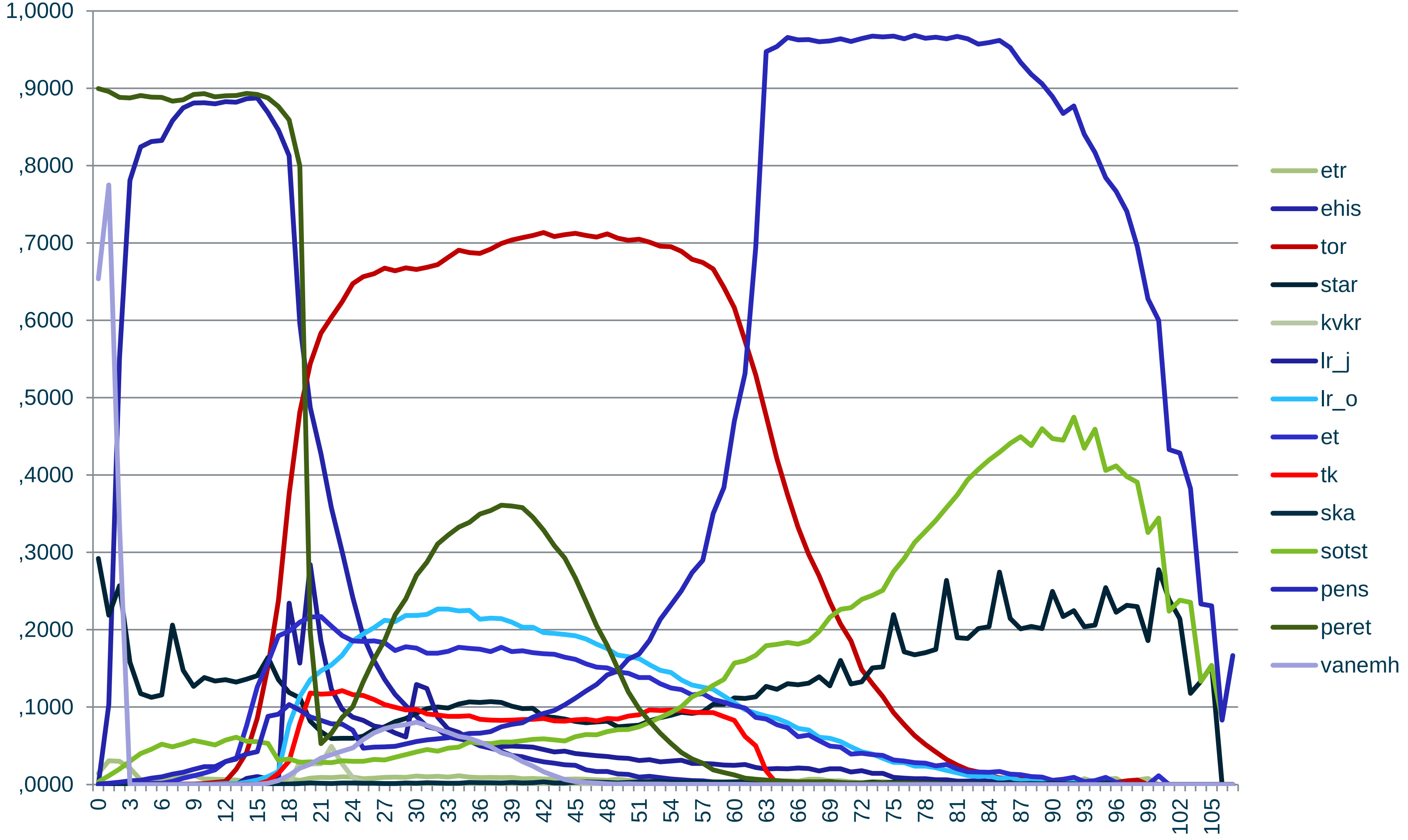
Total activity of R’s and NR’s in registers



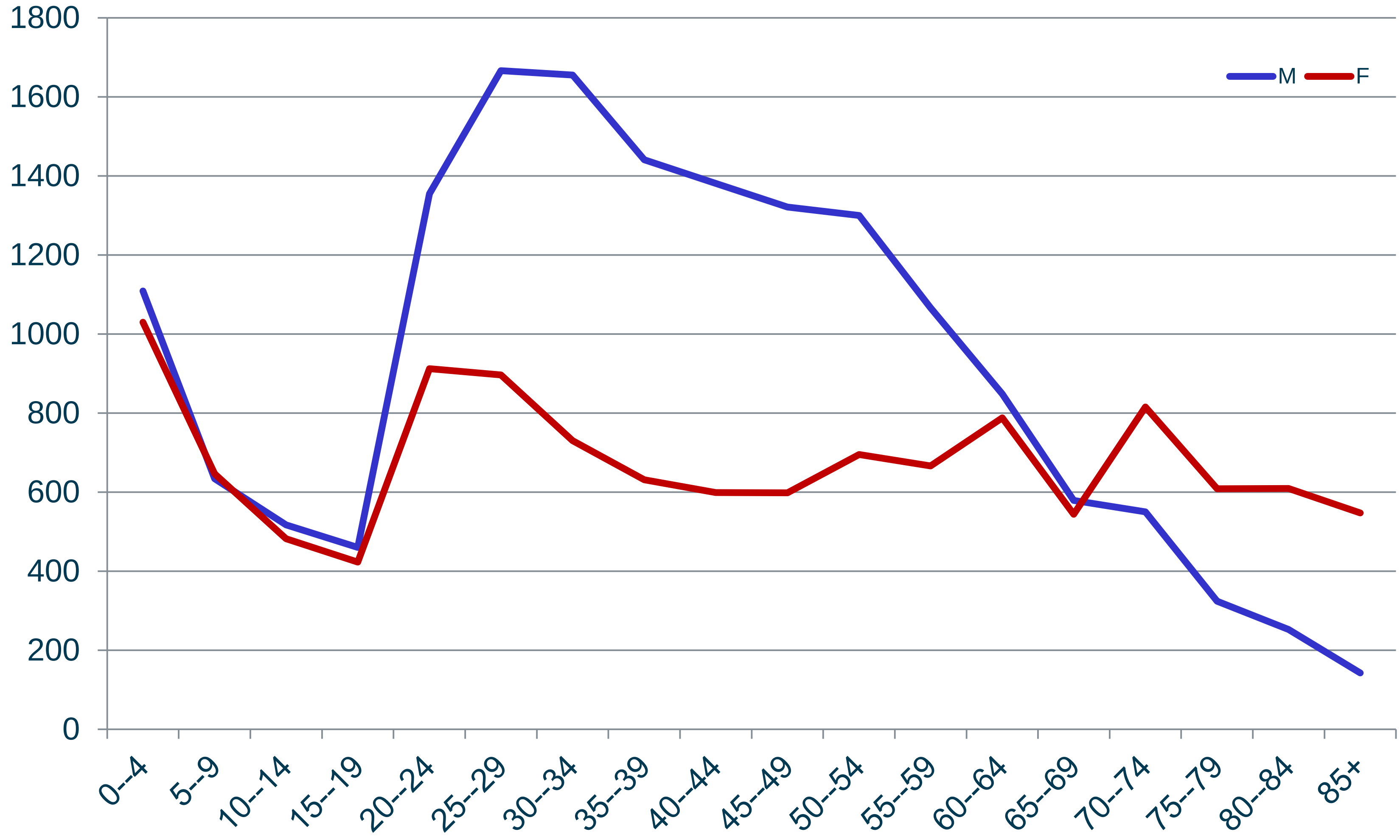
Residents’ activity in registers depending on sex



Distribution of activities in different registers by age



Age-sex distribution of estimated under-coverage (people added to census population)



eesti statistika

HOW TO GO ON?

The next steps

- The problem was solved, but this step initiated a series of new problems.
- The following census in Estonia is planned to carry through register-based. Which will be the census population this time?
- Usually, the basis for register-based census is population register.
- But there are many problems why the Estonian Population Register cannot be the exact basis of the following census:

Problems with PR’s population

- There might exist (long-time) inhabitants who have never registered as Estonian residents (under-coverage of PR);
- The process of non-registered emigration that caused the over-coverage of PR, detected after census 2011, might continue and cause increasing over-coverage of PR.
- Some people who have left Estonia without registering this step might have returned (again not registered). These people lessen the over-coverage of PR.
- Also there is possibility that registered (e-)migration does not happen in reality. Such occasions lessen the over-coverage of PR.

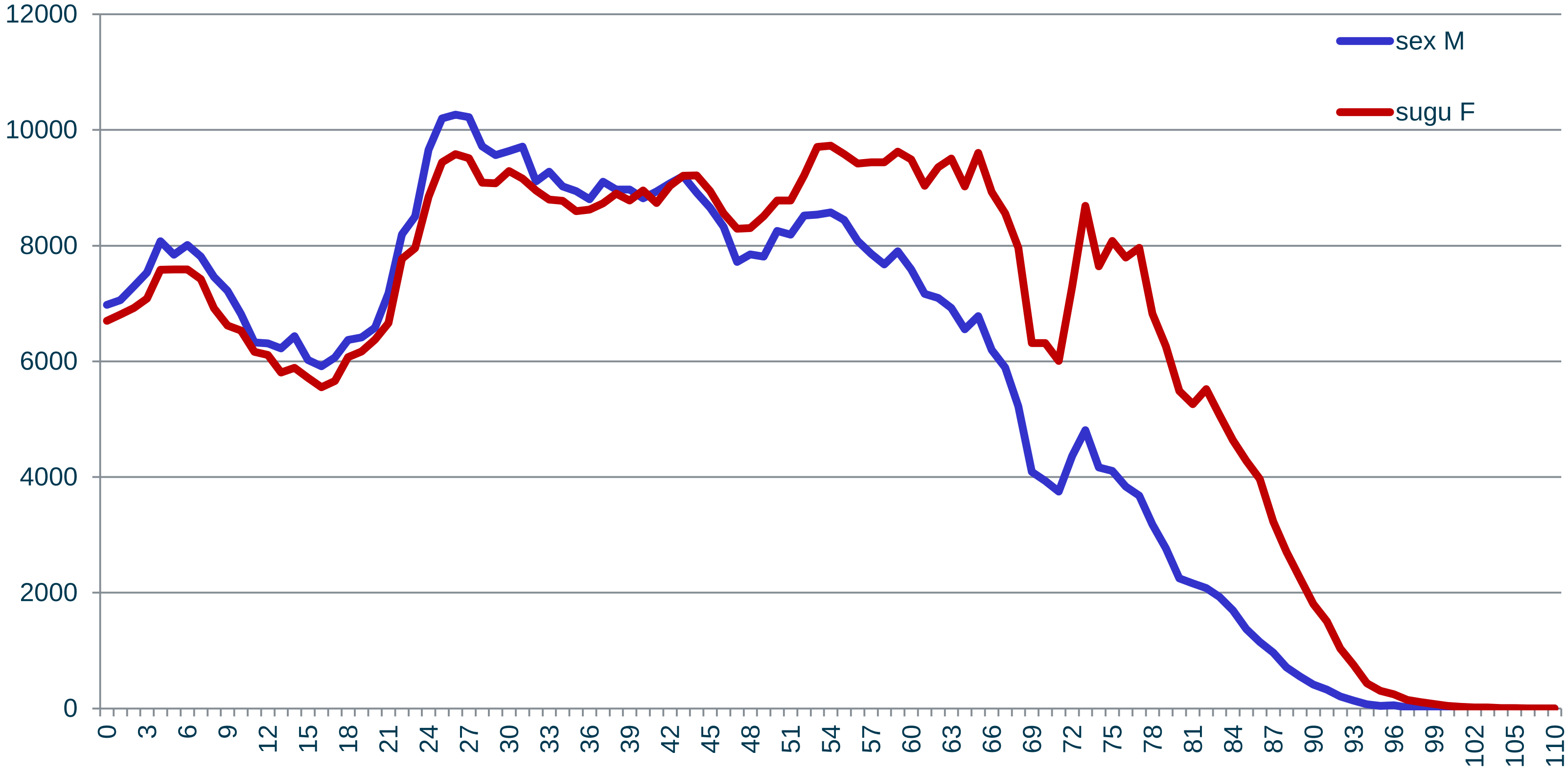
Problems with using census 2011 population as basis for following census

- The second possibility is to use as census population the census 2011 population that is corrected by natural increase and registered migration.
- In this case the same problems connected with registration of migration occur.
- Additionally, problems connected with census population arise:
  - ❑ Enumeration errors
  - ❑ Errors caused by estimated undercoverage.

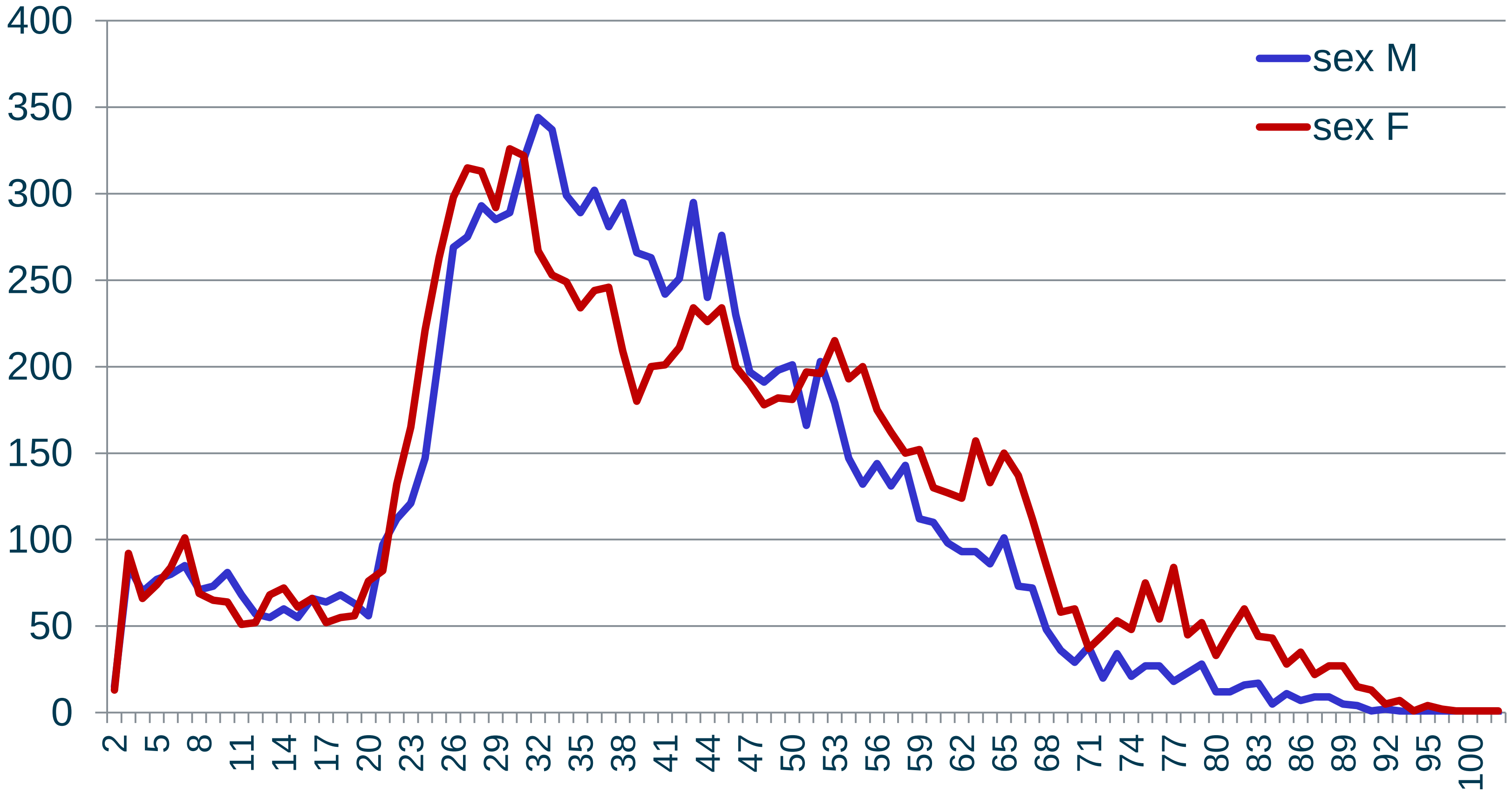
Population groups analysed in 2014/2015

1. Persons belonging to PR and active in at least one register in 2014 – 98,5% of official population size (OPS);
2. Persons not residents by PR, who have been active in registers 2014 – 1,8% of OPS
3. Persons who left Estonia after census (registered emigration; 0,7% of OPS);
4. Persons belonging to PR but not to census population (nor under-coverage; 2% OPS);
5. Persons belonging to PR and census population who were not active in any register in 2014 – 1,5% of OPS;

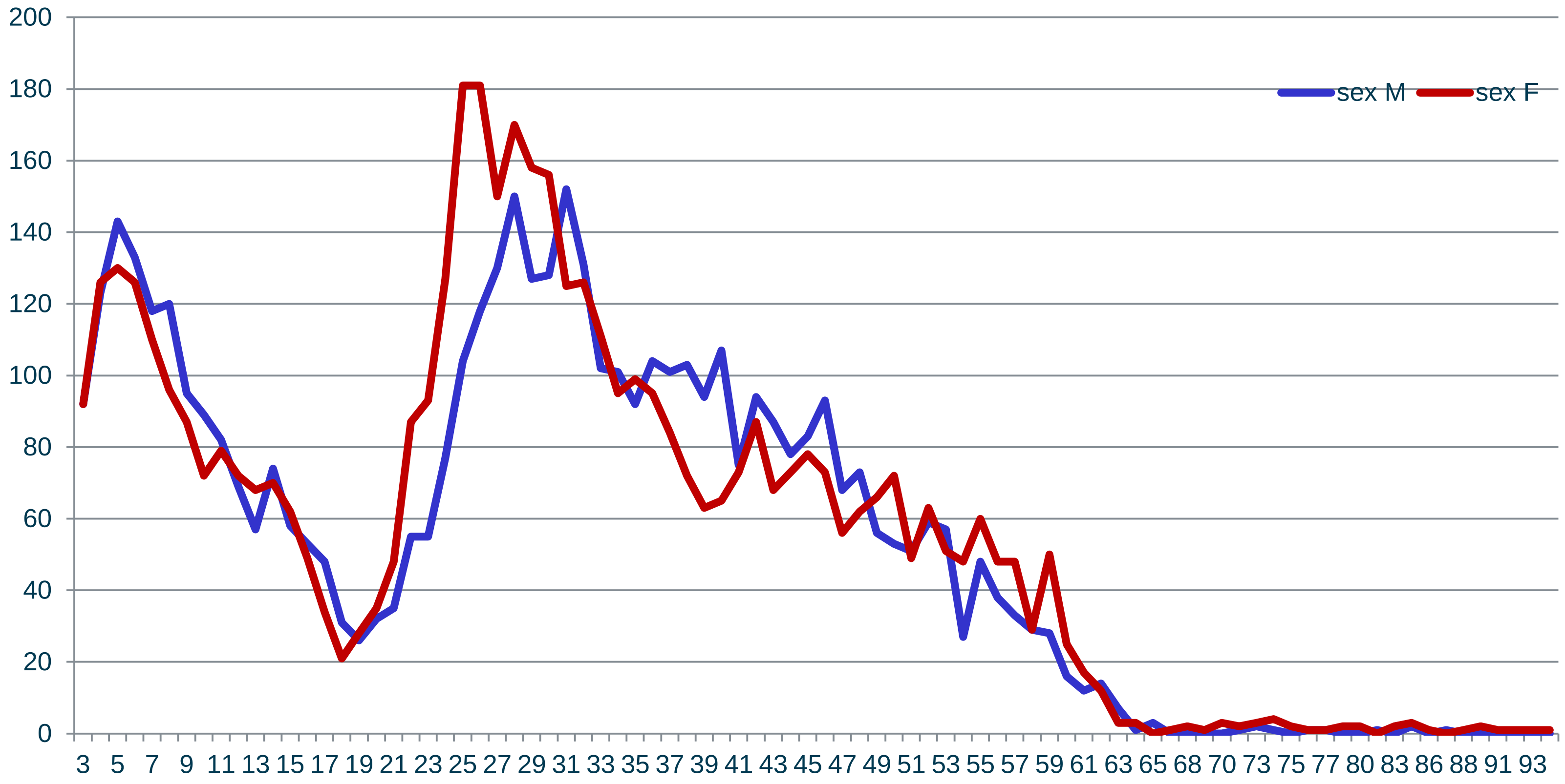
1. 1 300 000 persons belonging to PR population and having been active at least in one register in 2014 (98,4% of population).



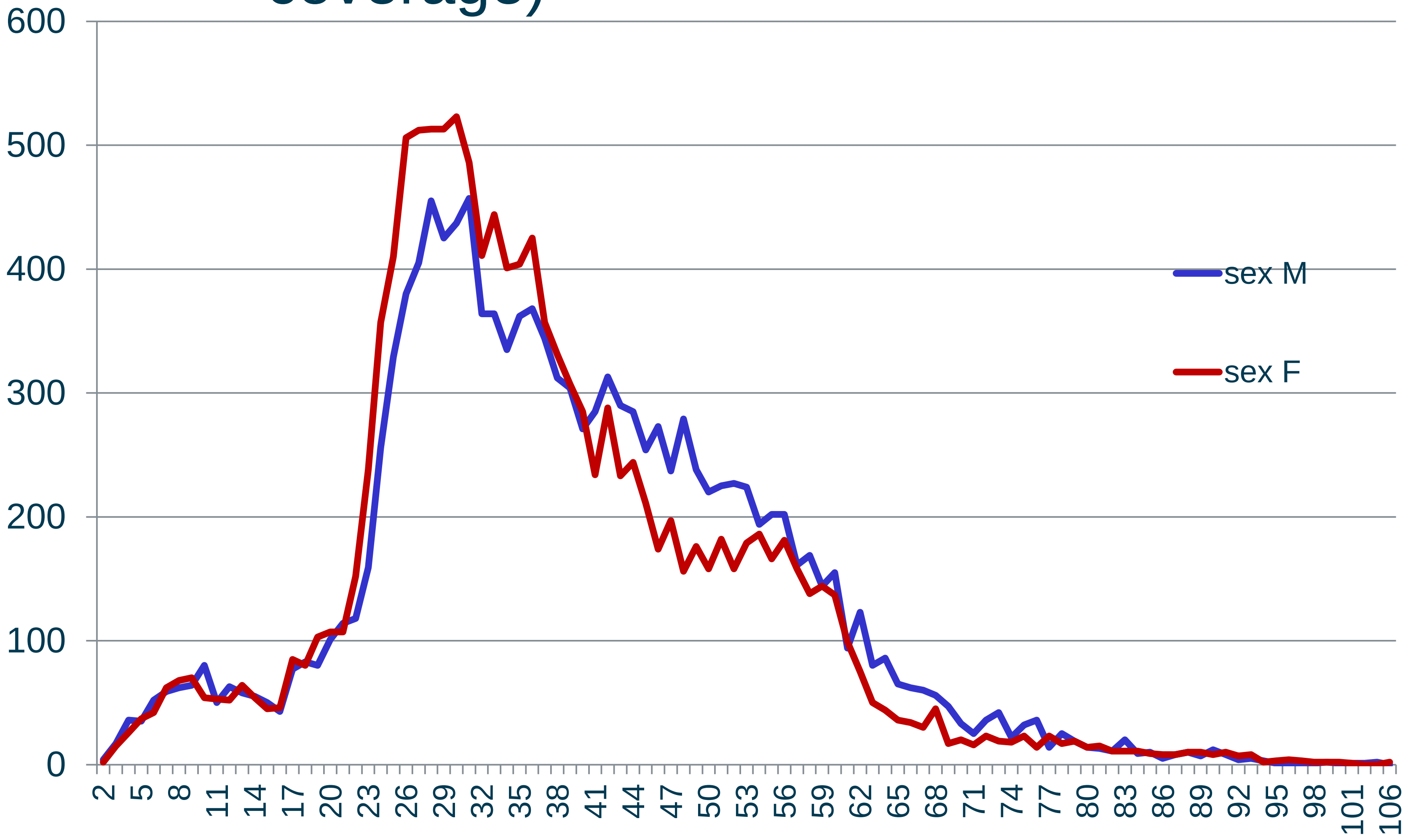
2. 23 114 persons – non-residents by PR, but have been active in registers 2014 (1,8% of population); 52% of them were in census population



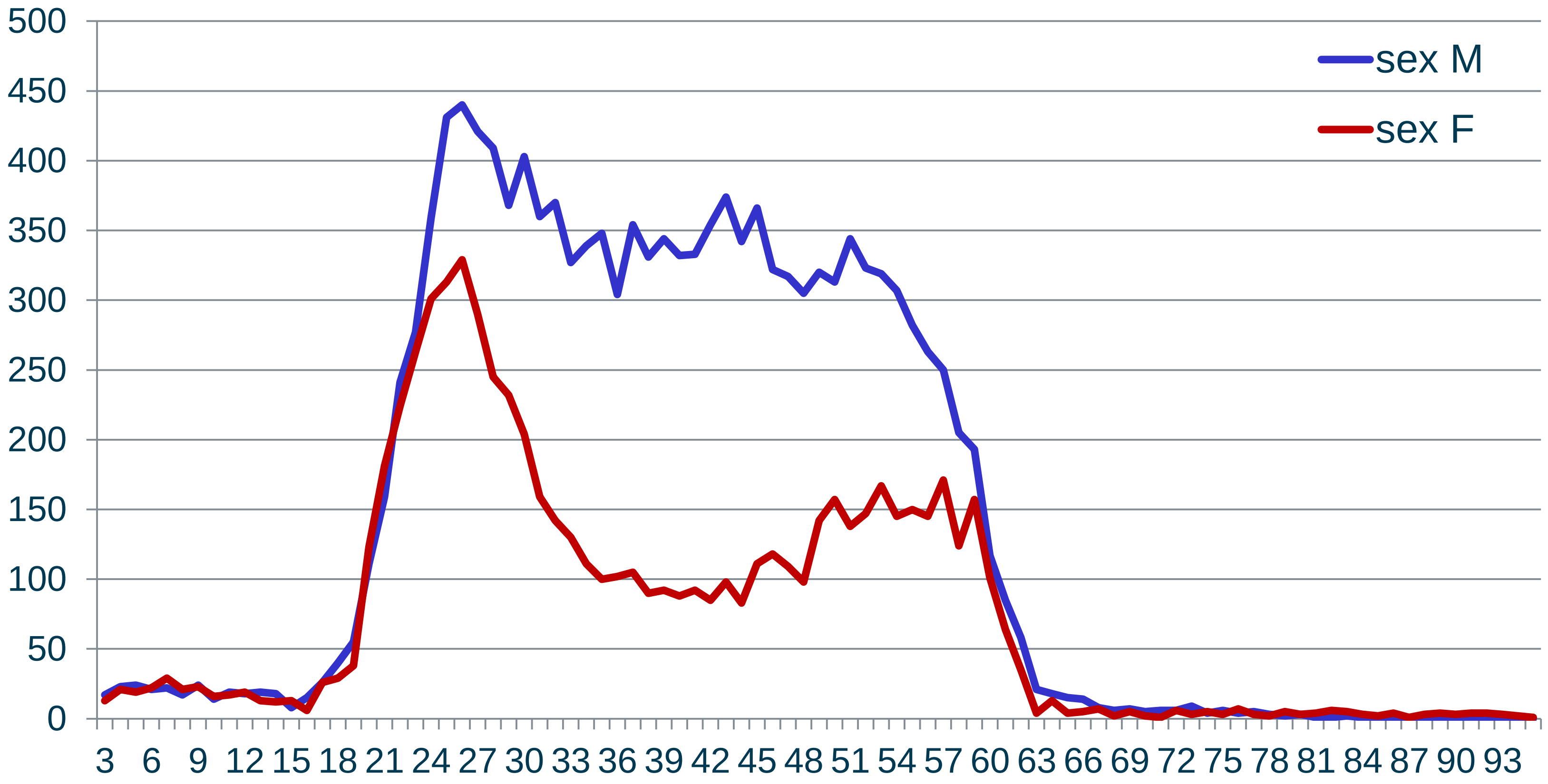
2. 9 596 persons who left Estonia after census (registered emigration; 0,7% of population); no activity in registers



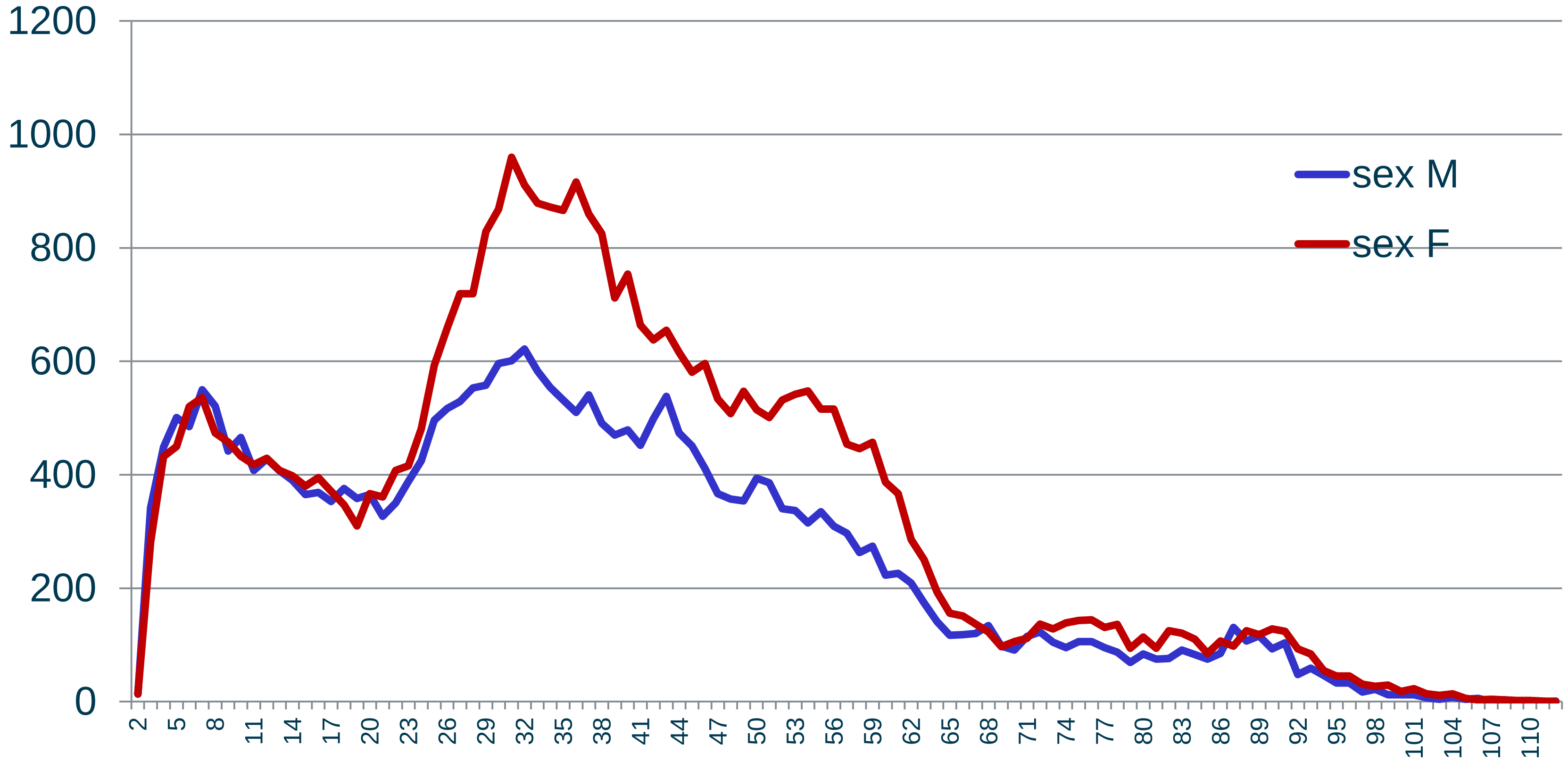
4. 26 542 people belonging to PR but not to census population (nor under-coverage)



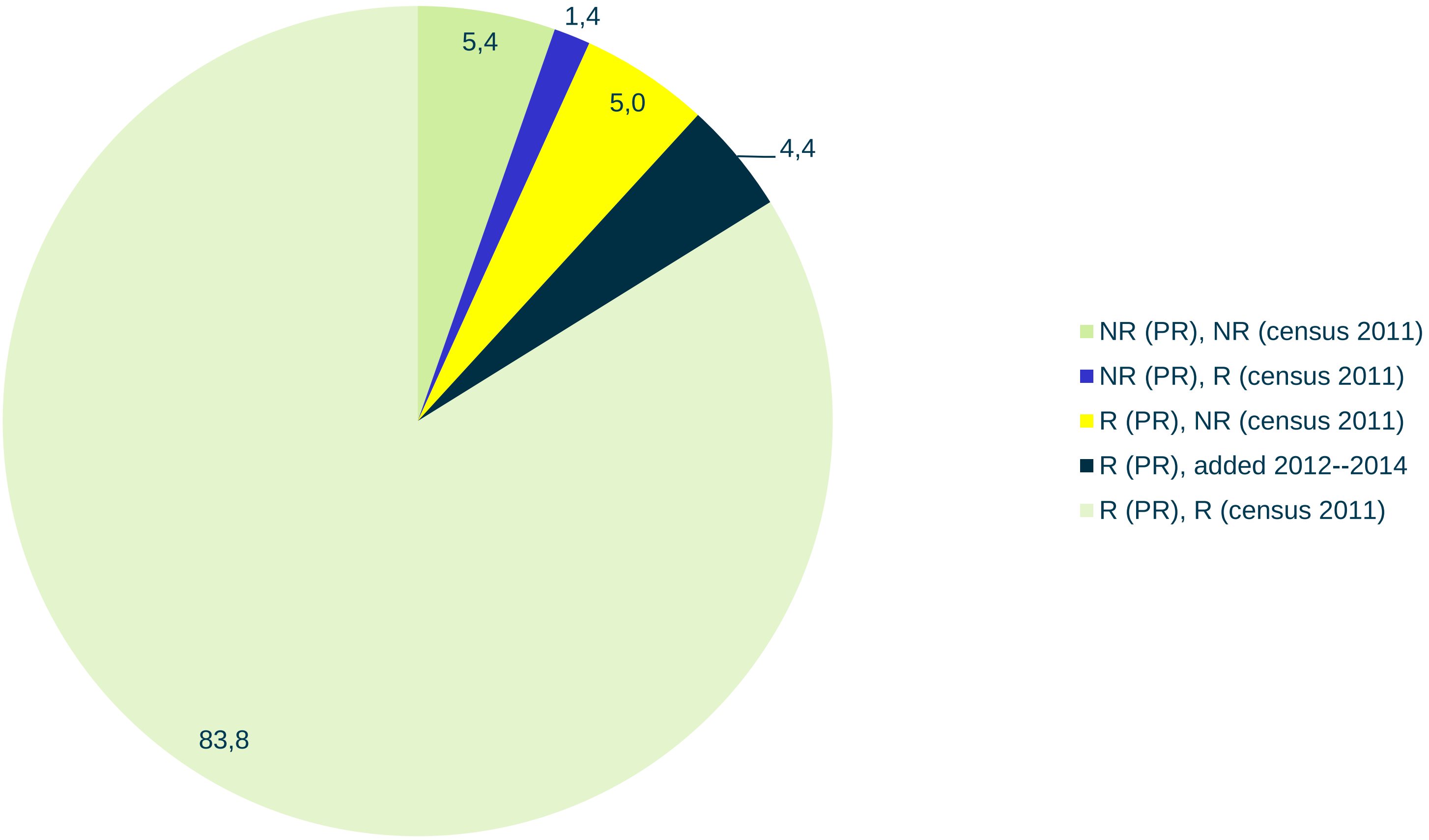
5. 20 576 people belonging to PR and census population who were not active in any register in 2014



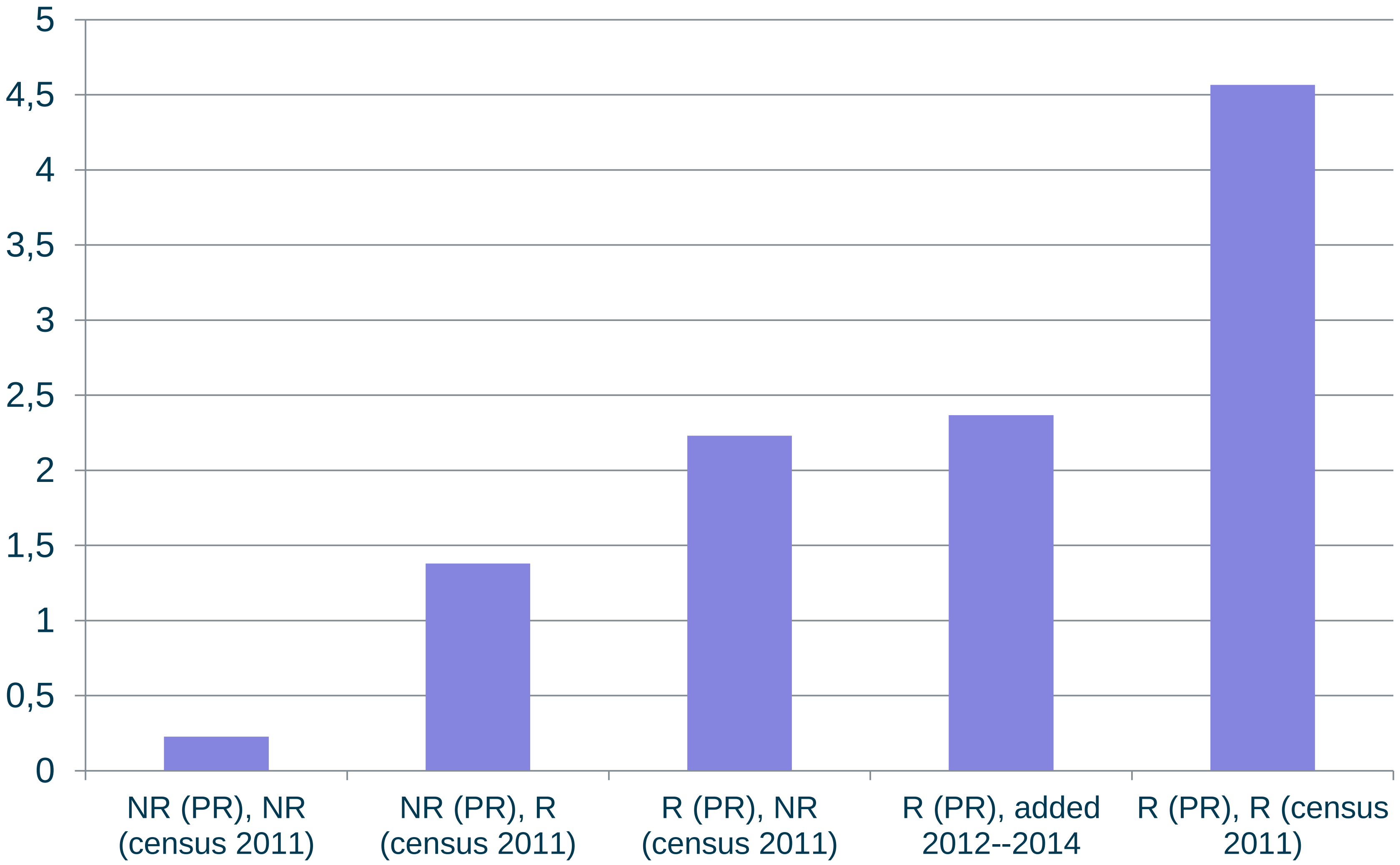
66 534 people who are NR by PR and do not belong to census population



Size of different population groups in PR (all groups together)



Average activity in registers of different population groups



RESIDENCY MODEL (ETHEL MAASING)

Residency model

- The most logical way to build the model was to use the same methodology used in estimating the overcoverage in 2012 [1—3]. But there were some complications:
- The population that should be tested is now bigger – consisting from all persons in PR;
- There is no good test-group, as it was in 2012, when the census data were available.
- But the list of registers was somewhat better than 3 years ago.

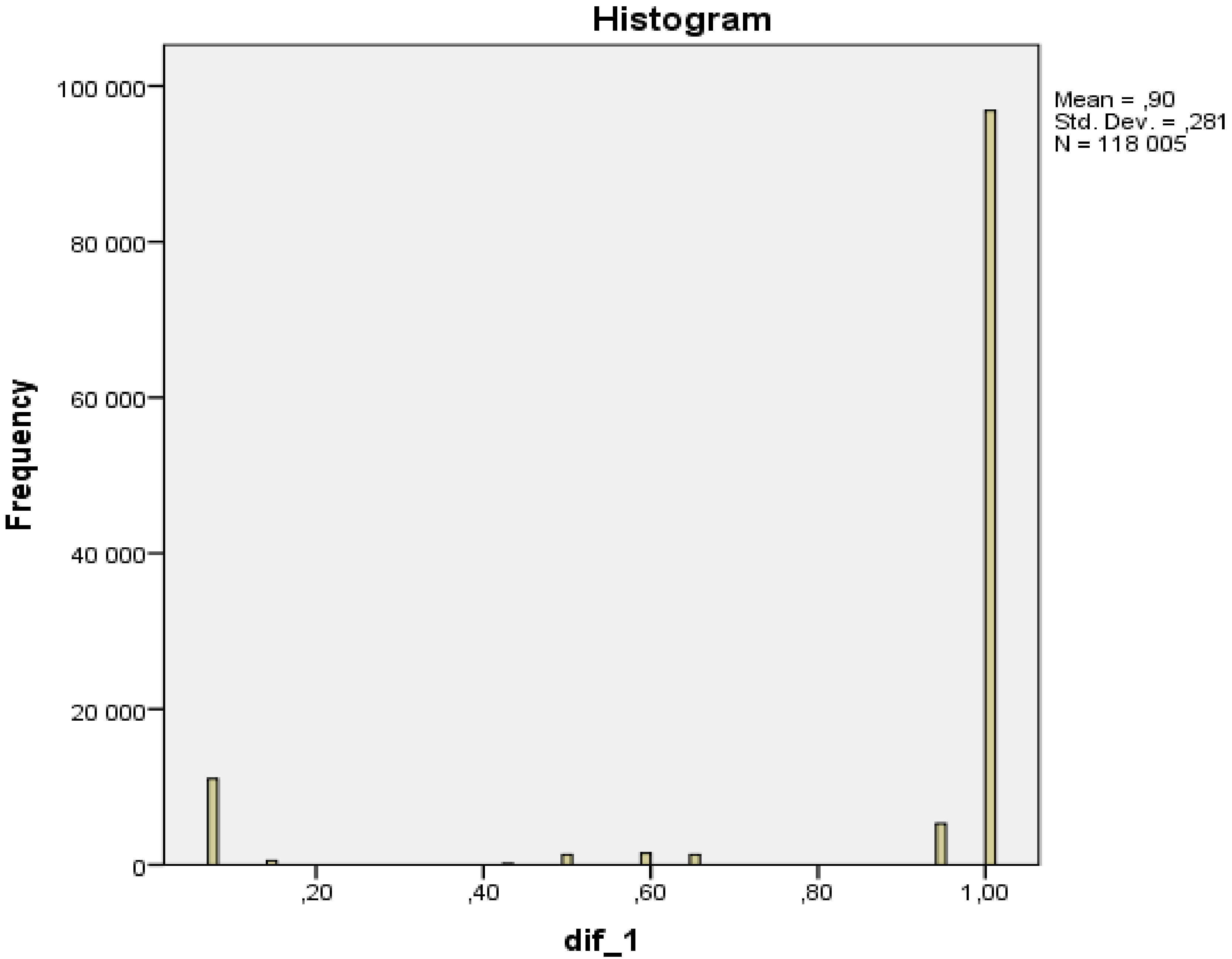
Model of Ethel Maasing

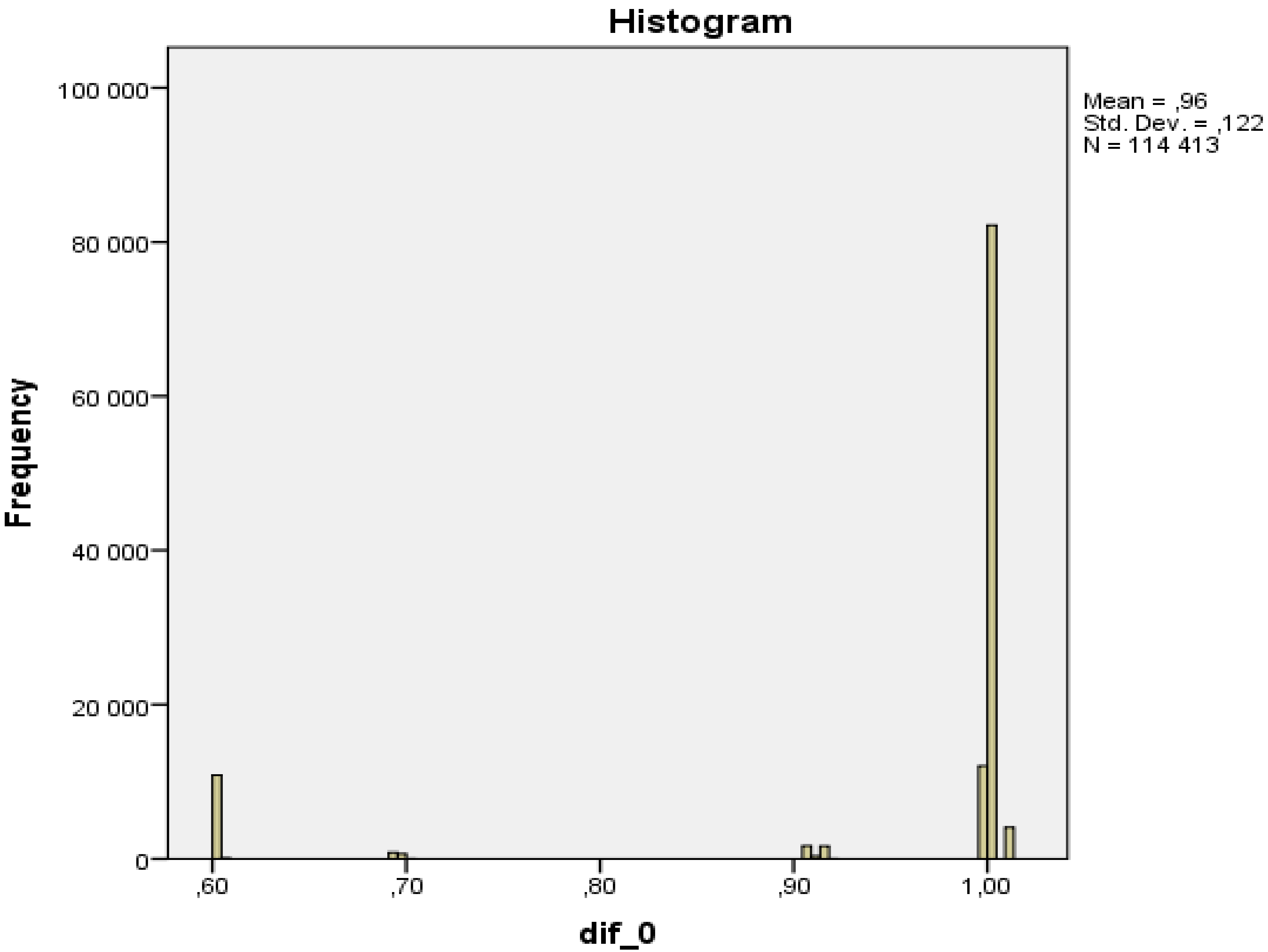
- Ethel Maasing solved the task to build a residency model in her master thesis.
- She used
  - more than 20 registers and sub-registers,
  - defined the test-groups using the information from census and PR,
  - used as methodology log-linear regression,
  - calculated threshold values in traditional way.

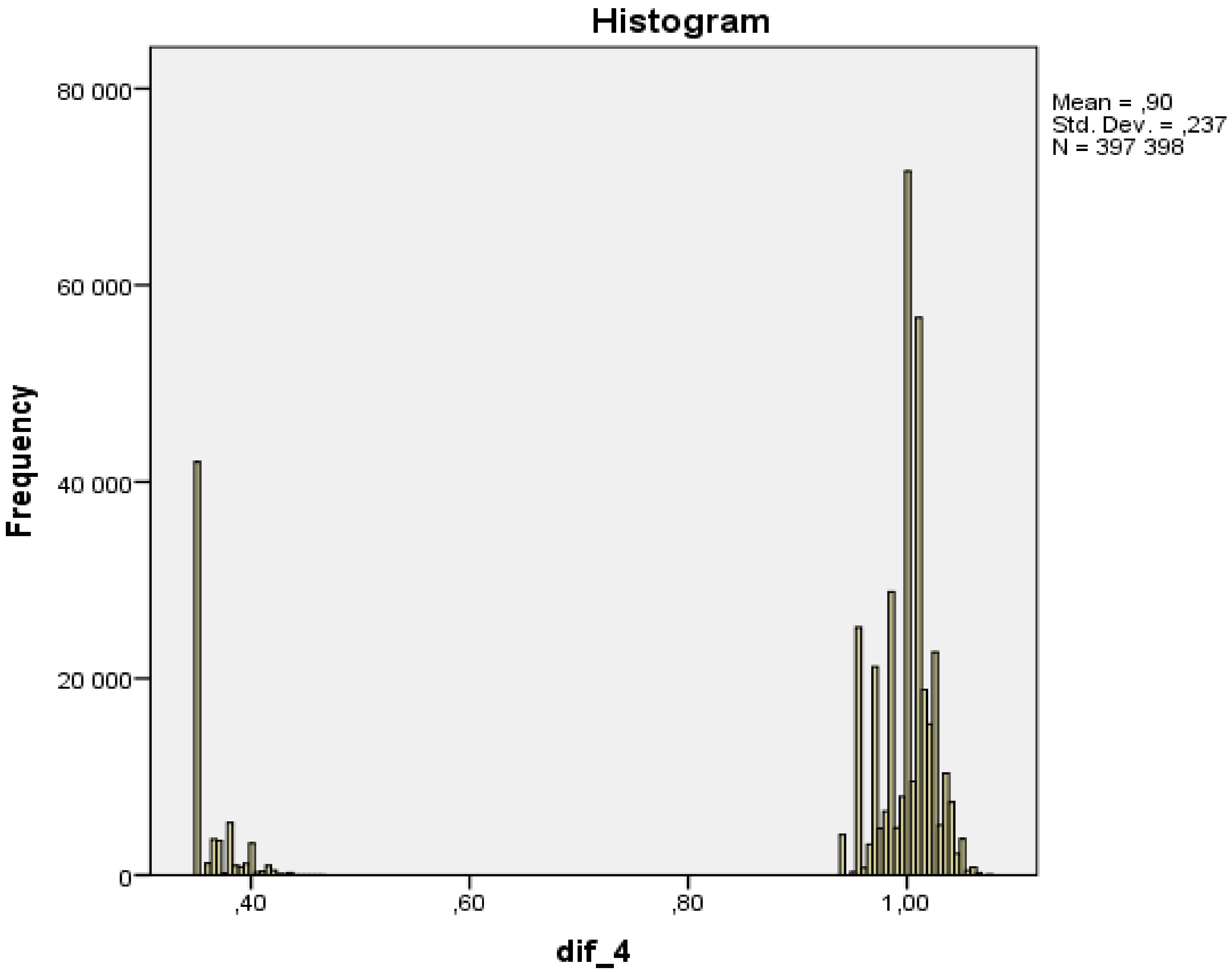
Ethel’s model gave quite good results, but the number of residents calculated by her was 1273 958, that is about 97% of official population size.

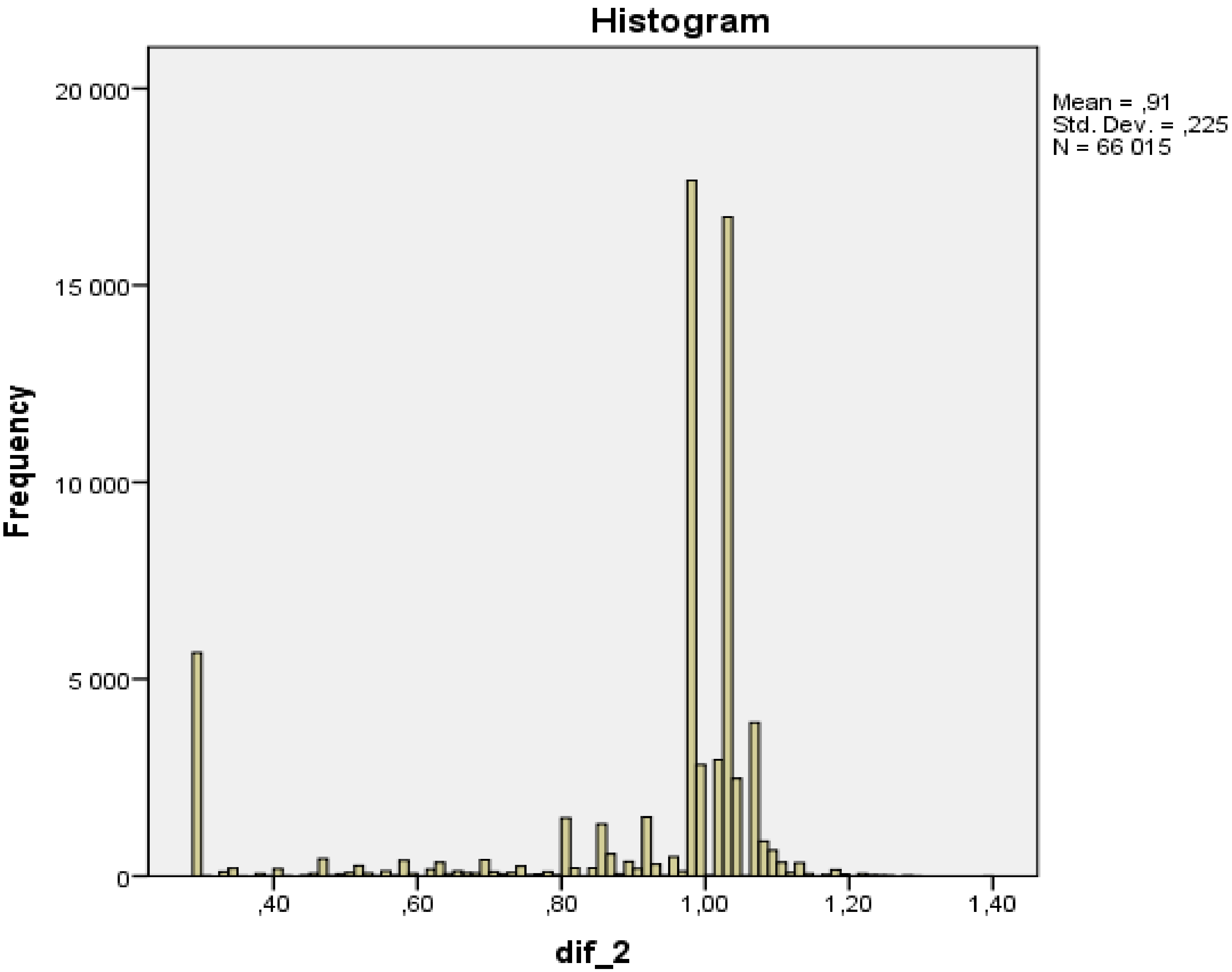
Seemingly, there were some residents who were by the model estimated to be NR.

- To specify the problem we look the histograms of model results.
- There are group of residents and group of non-residents, but also some intermediate people.
- Their differentiation depends on definition of threshold.

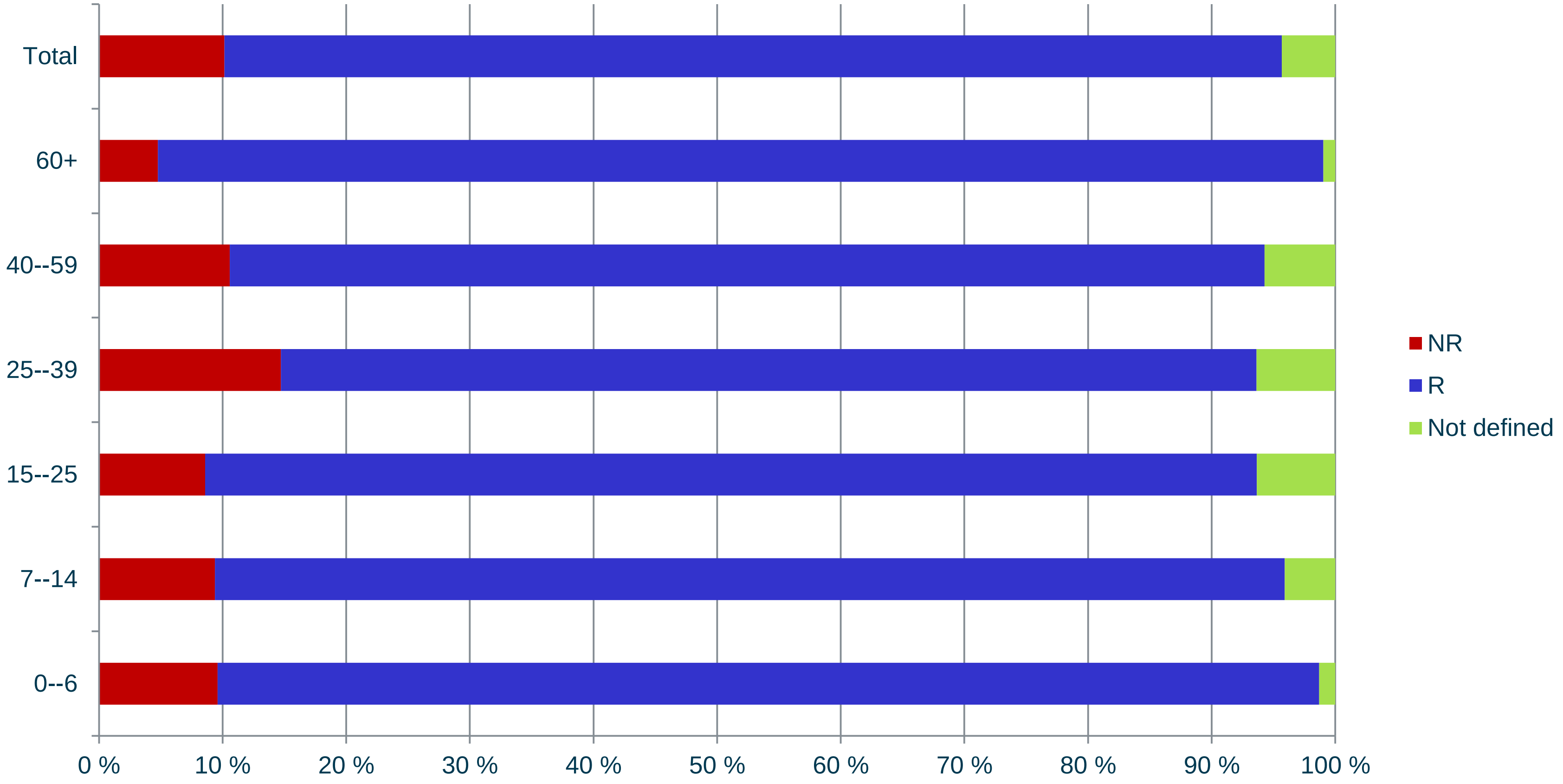

Histogram
Mean = ,90
Std. Dev. = ,281
N = 118 005
Frequency
100 000
80 000
60 000
40 000
20 000
0
,20
,40
,60
,80
1,00
dif_1


Histogram
Frequency
100 000
80 000
60 000
40 000
20 000
0
,60
,70
,80
,90
1,00
dif_0
Mean = ,96
Std. Dev. = ,122
N = 114 413


Histogram
Frequency
80 000
60 000
40 000
20 000
0
,40
,60
,80
1,00
dif_4
Mean = ,90
Std. Dev. = ,237
N = 397 398


Histogram
Mean = ,91
Std. Dev. = ,225
N = 66 015
Frequency
20 000
15 000
10 000
5 000
0
,40
,60
,80
1,00
1,20
1,40
dif_2

Model-based residency testing by age-groups. Not defined 4,3% of population



Why is some people's residency status „not defined"?

- The people whose residency status was „not defined" were active in some registers.
- These were not the right registers or not the most influential registers.
    - ❑ One possibility: these were registers that were often used also by NR (e.g. health insurance of children).
    - ❑ The second possibility: the people have a pattern of registers' activity that is different from the main group (e.g. young people in military service).

STABILITY OF SOLUTION

eesti statistika

Stability of residency algorithm

- In practice the people do not change very often their residency status. Also mobile young people rarely are commuting yearly between two or more countries, getting each case the different residency.
- From here it follows that also residency algorithms should have a stability property. That means, the part of population changing their residency status during a year should be restricted, for instance, their share must not be larger than a constant d (%).

Independence of model-based solutions

- The resident population defined by residence model for year n does not depend on the resident population defined for year n-1.
- The only source of dependence might be the same or partly same test-group.
- When using residency model, it is possible that big amount of people change their position near the threshold and hence also estimated residency status will alternate year by year.
- It makes sense to elaborate a methodology that avoids such instability and uses in defining person’s residency status for year n some information from his status in year n-1.

RESIDENCY INDEX

Residency index – an alternative methodology for testing residency

- Residency index is methodology for testing residency using peoples’ activities in registers (during a year), but this methodology differs from model-based methodologies by its continuity.
- The models just described in general do not use the residency status of persons in previous year(s), then residence index uses for defining residency status in year n substantially the status in several previous years (n-1, n-2 etc).
- This fact should guarantee the higher stability of residency decisisons.

One possible model for calculating the residency index RI(n)

- Let us have for each person an initial value of residency index RI(0) and the integrated residency activity for the same year S(0).
- The easiest way to calculate S(0) is to define for each register/sub-register a binary variable having values 0 and 1 depending on the activity of person in the register.
- The value of residency index for a person will be calculated by the following formula:

$$RI(n+1) = b \times RI(n) + a \times S(n),$$

where a and b are parameters to be estimated empirically.

The value of index is restricted by value 1:

$$\text{If } RI(n)>1, \text{ then } R(N):=1.$$

Interpretation of residency index as a probability

- As the value of RI(n) fills the condition

$$0 \leq RI(n) \leq 1,$$

it can be considered as an estimated (subjective) probability of a person to be resident.

For decision-making it is necessary to define a additional parameter – threshold c, so that

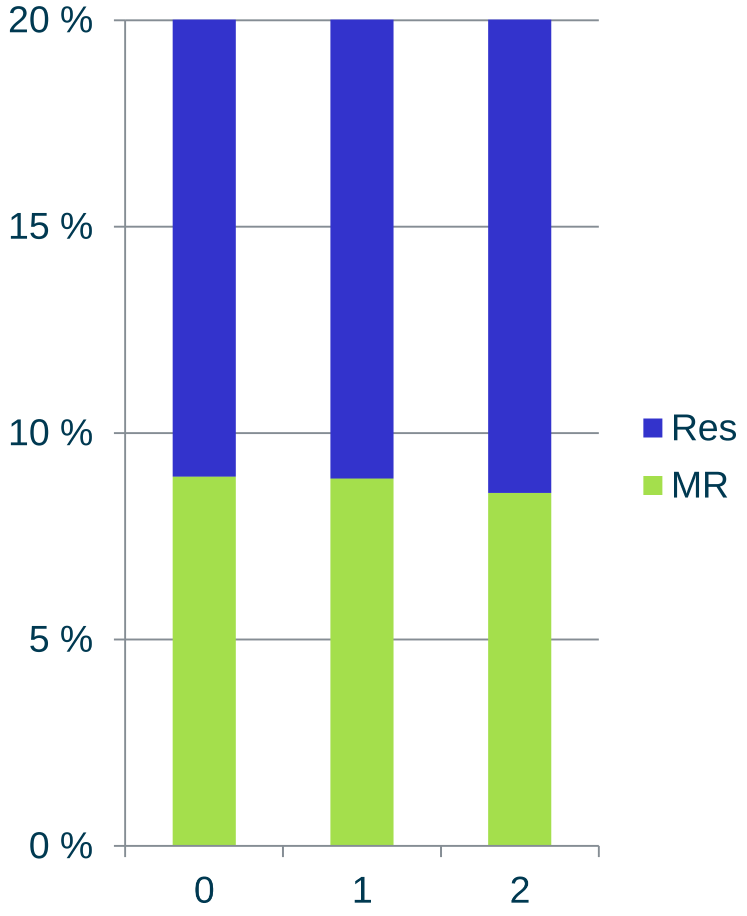
- If for a person in year n RI(n)≥c, then the person is estimated to be a resident,
- If RI(n)<c, then the person is estimated to be non-resident.

Possible modifications of RI(n)

- The parameter b is connected with stability of the algorithm – the larger is b, the bigger is the influence of previous year(s). If b =0, then the influence of earlier years is eliminated.
- Instead of the parameter a also the set of different parameters can be used.
- One possibility is to define these parameters as regression coefficients, but this makes the calculation much more complicated.
- Threshold value c should be taken so that both inclusion and exclusion errors are as small as possible. The values of errores can be estimated either empirically or also using simulation.

Example of calculating and using RI for years 2012—2014

- Initial values for RI(0)were estimated as 0, 0,4, 0,8 and 1 (depending on their belonging to PR, census population and registers.
- The parameters were taken: a =0,2, b = 0,8, c = 0,75.
- The sum of register activities was simulated by data of 2014.
- The results were quite close to calculated population sizes.



Comparison: residency model and residency index I

- For calculation of residency index it is not necessary to have test-data, that are difficult to get (if no censuses have been made recently).
- If the parameters are fixed, then in the case of using residency index decision can be made using the same algorithm for all persons (no need to use grouping the population to sex-age classes).
- In calculating residency index (the simplest way) all registers have the same weight, since persons, active in small registers can be classified to be R.
- Residency index guarantees considerably stabile values of solution.

Comparison: residency model and residency index II

- For using residency index the values of parameters a, b and c must be estimated and the quality of solution depends on these parameters.
- In calculating residency index (the simplest way) all registers have the same weight, since persons, active in non-differentiating registers can be classified to be R
- Residency index does not have standard procedure for error estimation.

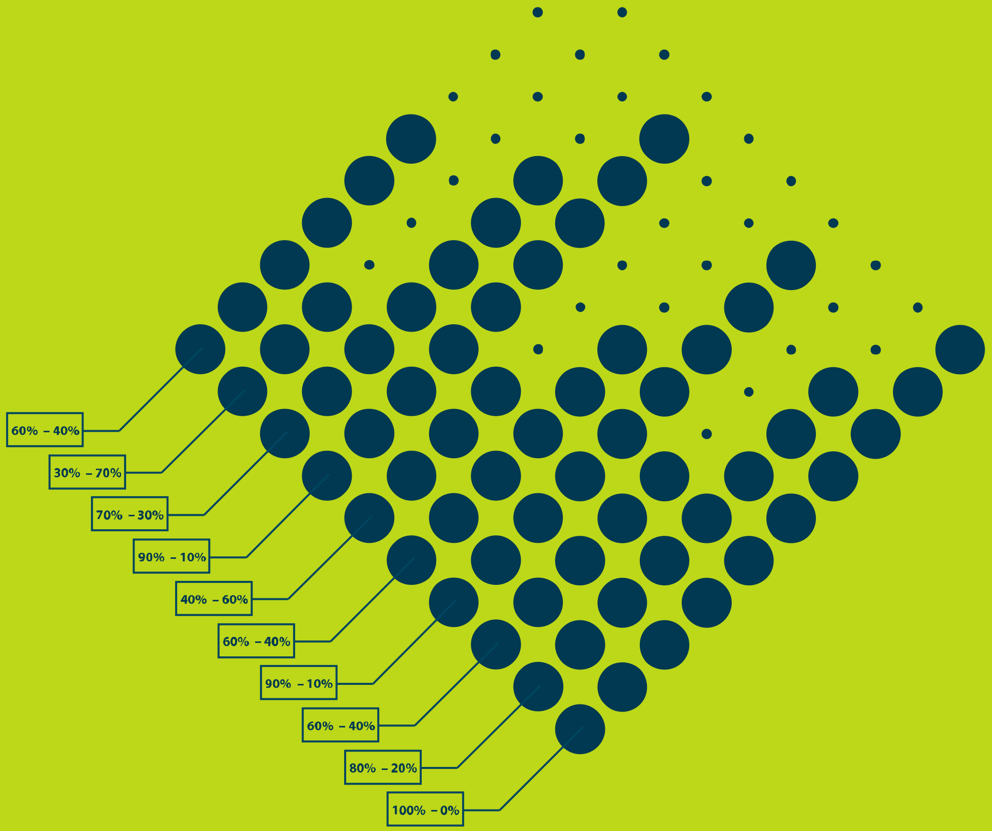
The initial data-source is for residency models and residency index the same – that is the set of data on activities of all population in all registers. The more complete the register data are, the better are both solutions.

Calculation

- Residency index is easy to calculate. The calculation can be made automatically using quite simple algorithm.
- But before using the algorithm the parameters a, b and c must be estimated.
- Probably, it will be useful periodically use some other model to check the results gained using residency index.

References

1. Tiit, E.-M. , Vähi, M. (2012) Enumerators' activity after the census. Quarterly Bulletin of Statistics Estonia, 2, 12 112--119
2. Tiit, E.-M., Meres, K., Vähi, M. (2012) Assessment of the target population of the census Quarterly Bulletin of Statistics Estonia, 3, 12 96—105
3. Tiit, E.-M. (2012) Assessment of under-coverage in the 2011 population and housing census. Quarterly Bulletin of Statistics Estonia, 4, 12, 116—119
4. Maasing, Ethel. Eesti alaliste elanike määratlemine registripõhises loenduses. http://dspace.utlib.ee/dspace;/handle/10062/47557



eesti statistika

Thank you for patience