United Nations ECE/CES/GE.41/2023/10

Economic and Social Council

Distr.: General
27 July 2023

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twenty-fifth Meeting
Geneva, 20–22 September 2023
Item 2 of the provisional agenda
Lessons learned from censuses of the 2020 round

Estonian 2021 census: determining households and dwellings from registers using graphs

Note by Statistics Estonia*[1]

Summary

The 2021 census was the first in Estonia to produce all EU-mandatory census characteristics from administrative data. A major challenge was the low, just 80% accuracy of the place of residence data in the Population Register and its impact on households.

In a register-based census, household is defined as a set of people living in the same dwelling. When using the place of residence from Population Register to determine households and families, the resulting statistics overestimates number of lone parents and underestimates number of couples.

To improve statistics on households and families, we have developed a graph-based method which uses input from administrative sources. We consider the people and addresses as nodes of a graph. A connection between two persons (such as marriage, parenthood) or a person and a place (such as real estate ownership) form the edges of the graph. A household is viewed as a subgraph containing household members and their dwelling. Then, determining households and their dwellings is equivalent to finding densely connected subgraphs, or in other words, to community detection.

To find connections between people or people and places we used data from 17 registers. Each edge in the graph was assigned a weight describing the probability of people living together or a person living on an address. The probability models were fitted on household data from existing surveys.

* Prepared by Helle Visk.

Note: The designations employed in this document do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

[1] This document was submitted late due to delayed submission of the paper by Statistics Estonia.

The new framework was used to compute households and place of residence in the census.

I. Introduction

1. The 2021 census was the first census in Estonia where all EU-mandatory variables were obtained from administrative data. Estonian register system is well-equipped for having a register-based census. Registers cover wide range of census topics. Also, linking sources is straightforward as we have unique identifiers for people, addresses and businesses.

2. Although data quality in registers is generally high, there are some exceptions. For example, the place of residence data in Population Register (PR) is accurate in only about 80% of people (Gortfelder & Puur, 2021). The reasons why people do not keep their information in register up to date include considering registration unnecessary, using services and benefits of certain municipality, perceiving current home as temporary, among others (Gortfelder & Puur, 2021; Äär, 2017).

3. In a register-based census, a household is composed of people living on the same address, regardless of having common budget or not. Family is defined in a narrow sense, as a family nucleus. It is either cohabiting or married couple with or without children, or lone parent with one or more children. Family consists of people living in the same household. Since place of residence is the basis for dividing population into households, its inaccuracy affects household and family statistics.

4. Statistics Estonia conducted a pilot census in 2016, the households were derived from place of residence data from PR. The household and family statistics differed substantially from 2011 census. For example, the number of lone parents was 67% higher compared to 2011 census, number of partners was 26% lower.

5. Overestimating number of lone parents is characteristic to situation where family members are registered on different addresses. For instance, let us consider family of four: mother, father, daughter, and son. If father and daughter register themselves on a different address, they will appear as two lone-parent families (mother-son, father-daughter) in PR.

6. As households and family statistics derived from PR were heavily biased, the place of residence data could not be used for census per se. To obtain better statistics, we needed to find a method for reuniting families that appeared broken in the registers.

II. Methods

A. Households or dwellings

7. Reconstructing families is more than identifying links between possible family members. Let us consider the example from paragraph 5. Reuniting this family means that father and daughter “move back in” to mother’s and son’s home. Note that this involves assigning father and daughter to a dwelling that is different from the one observed in PR. Also, the household composition would be different from PR for all family members. Hence, reuniting families implies that household composition and dwellings would change.

8. Information to determine households and their dwellings can be gathered from administrative sources. For example, data on parenthood and marriages helps to identify people in the same family. Data on property and electricity contracts may be useful to find homes of people.

9. It is not clear whether it is better to start with determining the households and then assigning each household a dwelling, or first find homes for each resident – then households form of people on the same address. If starting with households we fail to find families that have no direct ties, like cohabiting partners who co-own an apartment. On the other hand, if

people are linked to addresses while ignoring data that connects people (e.g., marriage) it is unlikely that the resulting families are much more realistic than in PR.

B. Network of people and places

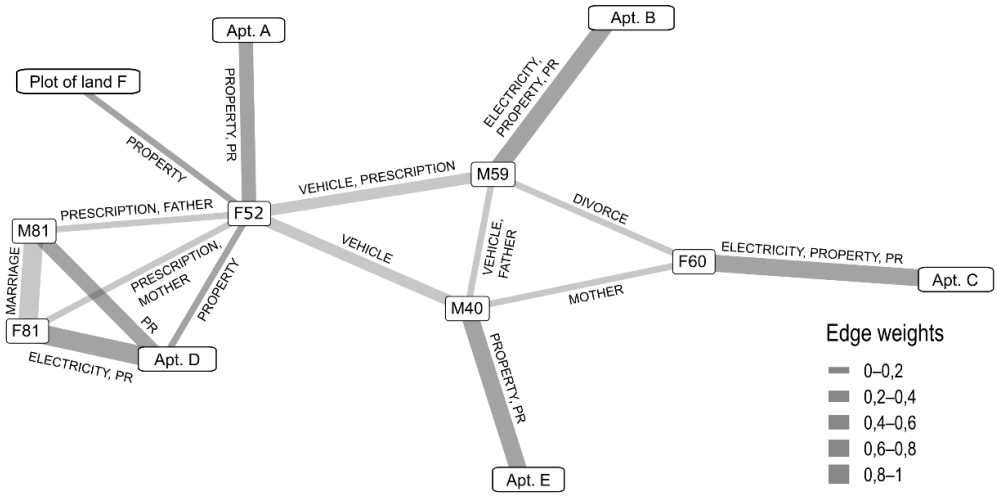
10. Our idea is to simultaneously construct households and assign dwellings. For that, we consider people and dwellings as nodes of a graph. The edges of this graph are given by:

(a) Connections between people (such as marriage, parenthood, buying prescription drug for someone, sharing a car); or

(b) People and places (e.g., place of residence, property, electricity contract).

A sample of this graph is shown on Figure 1.

Figure 1
A fragment of the graph of people and places



Note: The nodes are people (label shows sex and age) and places (apartments and one plot of land). The edges join people (lighter grey) or people with places (darker grey). Edge weights show probability of people living in the same household or person living on the address. The labels of the edges show type of relationship. ELECTRICITY: person has electricity contract on given address, VEHICLE: people are linked to the same vehicle as owners or users, PRESCRIPTION: person has purchased prescription drug for the other, PR: place of residence in Population Register. The figure has been previously published (Tiit et al., 2021).

11. We are interested in finding division of the graph of all people and dwellings into subgraphs, each containing members of one household and their dwelling. It is natural to assume that people in the same household are connected strongly and people have strong connection with their home.

12. In graph theory, a community is defined as "group of nodes that are relatively densely connected to each other but sparsely connected to other dense groups in the network" (Porter et al., 2009). We conclude that finding households and dwellings is like finding communities. Community detection is a common task in the analysis of networks and many algorithms are available in various programming languages.

13. Links between people, or people and dwellings, vary in strength. For example, almost all underaged children live with their parent(s), but it is less common once the child becomes adult. Hence, it makes sense to assign weights to edges.

C. Workflow

14. First step is to collect the data. We call different types of connections person-to-person or person-to-place signs. The data on signs was acquired from 17 registers (Table 1, Table 2). Additionally, data from large annual household surveys – Estonian Social Survey / EU Statistics on Income and Living Conditions (EU-SILC) 2021, Estonian Labour Force Survey (LFS) 2021 – were used as training data to model relationship between signs and real-life patterns.

Table 1
Person-person signs from the registers

Register	*Person-to-person signs*
E-File	Persons are on the same side of an alimony dispute
	Persons are on the opposite sides of an alimony dispute
Health Insurance Information System	One person has cared for another person in the year preceding the census
Traffic register	Persons are linked to the same vehicle (e.g., user and owner of the car)
Register of taxable persons	Persons have co-applied for mortgage
	One spouse has transferred tax-free income to the other spouse
	Using income tax benefit for two and more children (link is between the child and person submitting the declaration)
	Person using income tax benefit for educational expenses of another person
Estonian Medical Prescription Centre	Person has purchased other person’s prescription drugs
Population register	Persons are married
	Persons are divorced
	An adult serves as a guardian for another adult
	Person is the mother of the other
	Person is the father of the other
	Person has full right of custody over child
	Person has limited right of custody over child
	Child is separated from parent
Social Services and Benefits Registry	Persons have received subsistence benefit in the same household
Social Security Information System	Person receives family allowance for a child
	Person receives parental benefit for a child
	An adult receives extra leave to care for a disabled adult

Table 2
Person-place signs from registers

	Register	*Person-to-place signs*
Potential dwellings	Elering (electricity system operator)	Person has an electricity contract at the address
	Register of persons registered as unemployed or jobseekers, and of provision of labour market services	Person's place of residence
		Person's postal address
	Prisoners' register	Place of residence of probationers
	Land register	Real estate belonging to person
	Population register	Person's registered place of residence
		Person's additional address
		Person's previous places of residence
		Person's place of stay (e.g., dormitory)
	Population and housing census of 2011	Addresses of the person and his or her mother
	Social Services and Benefits Registry	Person's place of residence
	Register of taxable persons	Real estate purchased with a person's housing loan
Municipality level	Estonian Education Information System	Kindergarten of child
		University or vocational school student
		School of pupil in general education
		Teacher's place of work
	Health Insurance Information System	Dental care institution visited by the person
		Medical institution visited by the person
		Person's family physician (GP)
	Identity Documents Database	Place of receipt of an identity document
	Mandatory Funded Pension Register	Address of person who has joined the funded pension system
	Estonian Medical Prescription Centre	Pharmacy in which the person has purchased medication
	Employment register	Person's place of work

15. Edge weights were modelled as:

(a) Probability of people living in the same household (logistic regression model was fitted); or

(b) Probability of person living in a dwelling (random forest). The latter model also included municipality level data (such as having general practitioner in certain municipality), and distances from kindergarten, school, and work.

The models were fitted on survey data and then applied to whole population.

16. The graph included in total 5.2 million nodes and 7.8 million edges. The community detection was applied in two phases:

(a) First, Louvain method (Blondel et al., 2008) was used to break the initial graph to subgraphs of up to 5000 nodes;

(b) On each of those subgraphs, Infomap algorithm (Rosvall & Bergstrom, 2008) was applied recursively until the communities were small enough or the modularity did not improve significantly.

17. The resulting communities were similar to households and the family statistics improved compared to the PR (Table 3, Figure 2). However, numbers of multifamily households and families with adult children were overestimated, number of single person households was underestimated. Also, small number of children were in households without adults.

Table 3
Distribution of family status of people in the training data

Family status	*Training data*	*PR*	*Clusters*
Partners, married	34.6	27.6	34.6
Partners, cohabiting	16.3	10.4	14.5
Lone parents	4.2	9.6	4.8
Child, not of lone parent	24.2	19.6	27.3
Child, of lone parent	5.6	13.5	5.9
Not in a family nucleus	15.1	19.3	12.9

Note: Training data consists of people from EU-SILC and LFS 2021.

Table includes data of 32,802 persons who were present in all sources (this excludes non-residents, members of institutional households on census moment 31 December 2021 00.00, people born after this moment, and those who died before). The data is unweighted..

18. In postprocessing, children living alone were added to households of their parents or some other related adult. Using survey data, heuristics were developed to break some of the least connected communities to smaller parts (e.g., if there were multiple families in some household, we considered separating the least connected family to a different household).

19. In the next phase, each household was assigned a dwelling. This task is trivial if there is only one dwelling in the community. However, there were communities with multiple dwellings to choose from, and some communities did not have any. Also, after postprocessing described in paragraph 18 there were communities with multiple households competing for the available dwelling(s). Generally, each community included $m \geq 0$ dwellings and $n \geq 1$ households.

(a) In each community, strength of connection between households and dwellings was derived from person-dwelling weights. We preferred household-dwelling combinations that were most strongly connected. In case of ties, we gave priority to dwellings with higher electricity consumption and larger dwellings. In this step, 96% of households were assigned a dwelling:

(i) For the rest of the households, most probable municipality was selected. Also, we computed an anchor point for each household in their selected municipality based on the geographical coordinates of places the household members had connections with;

(ii) Home candidates for this step were selected from dwellings that were left unoccupied. We considered places that household members have connections with,

either directly or via other people. An example of potential home could be apartment of household member's mother. For each household, we only considered dwellings that were in their selected municipality;

(iii) There was many-to-many relationship between households and dwellings. Some households had multiple reasonable candidates for home. On the other hand, some dwellings were candidates for different households. In such a case, there are many ways how to match households and dwellings. We considered the households and dwellings as a bipartite graph and opted for stable matching. In a stable matching, there is no household-dwelling combination that would prefer each other over their matched "companions" (Gale & Shapley, 1962);

(iv) A prerequisite to compute stable matching is a sort of ranking: which households prefer which dwellings and vice versa. We declared that households prefer dwellings that are 1) close to their anchor point and 2) larger. Dwellings preferred 1) larger households, 2) closer households;

After this step, 99.4% of households had home;

(b) Other households were given a random unoccupied dwelling close to their anchor point.

20. People in the institutional households and homeless people were handled separately. Lists of homeless people were provided by municipalities; members of institutional households were known from registers.

III. Results

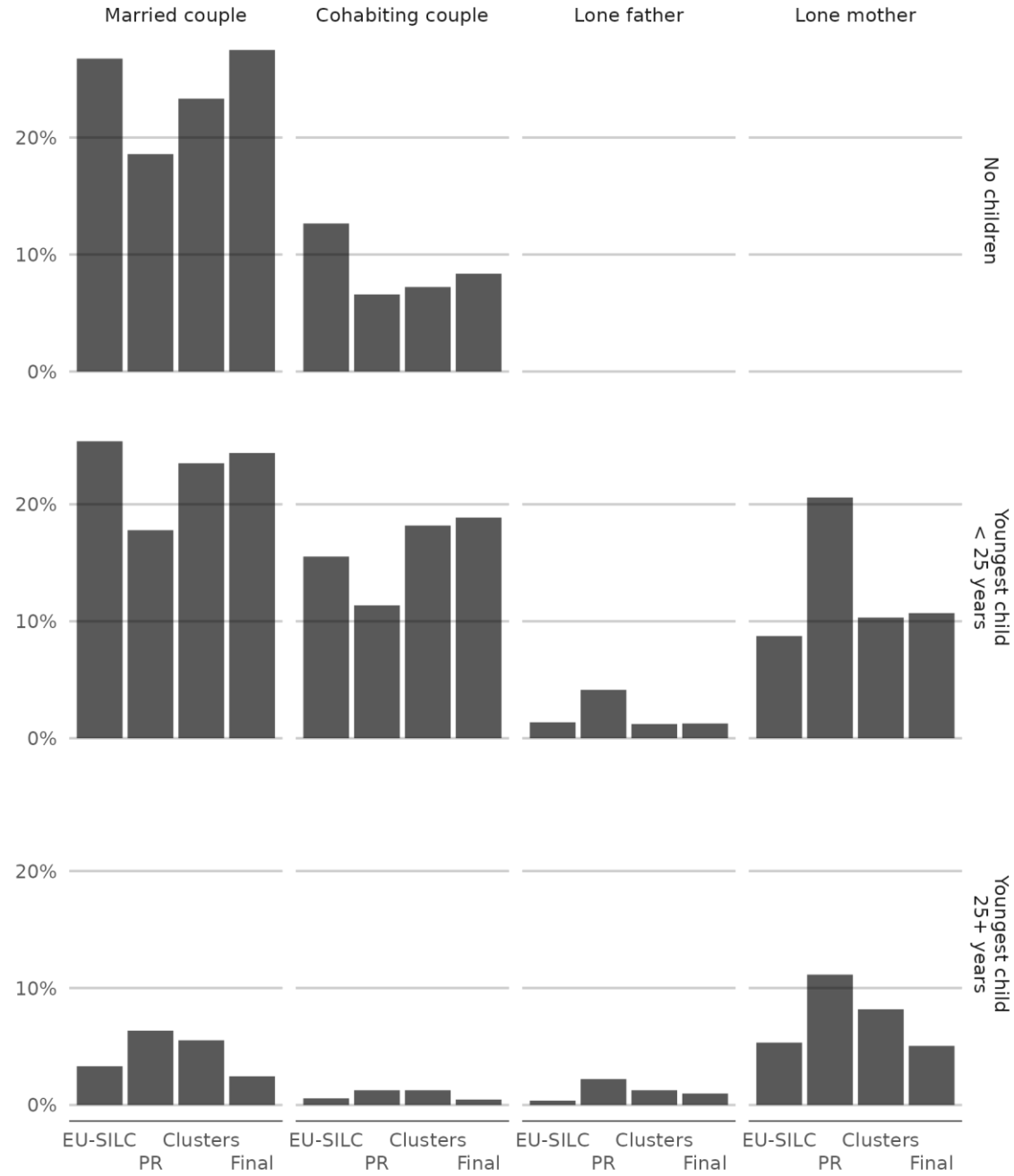
21. The families and households' statistics for the census were calculated using the new method. It also replaced PR as the basis for the geographical breakdown of yearly population statistics starting from 1 January 2022.

22. Despite using an array of sources for dwellings, 3 people out of 4 retained their place of residence as it was registered in PR. As some people were assigned different dwellings, the population of municipalities changed compared to PR. Out of 79 municipalities, 35 remained roughly as populous as in PR (±2%), 23 were at least 2% smaller, and 21 gained more than 2% of population. The losses were greatest in small islands (Ruhnu -29%, Vormsi -23%, Kihnu -19%, etc.) and other popular areas for summer homes (Alutaguse -8%, Narva-Jõesuu -6%). The top winners were Russian-speaking towns from northern Estonia (Loksa +7%, Maardu +6%, Kohtla-Järve +4%).

23. The motivation behind the graph-based approach was to improve household and family statistics. In the Figure 2, we explore the distribution of type of family nucleus according to different sources. As a benchmark we use the data from the 2022 European Union statistics on income and living conditions (EU-SILC 2022), collected 1 to 5 months after the census moment. This data is the basis for yearly household and family statistics. The PR data shows families derived from the registered place of residence. As in the pilot census, the share of families of partners is underestimated and lone parent families are inflated in PR. By applying community detection, the number of each type of couples grows, although the rise is modest in cohabiting couples without children. Finally, after postprocessing, we get distribution that aligns well with EU-SILC data. We still observe mismatch among cohabiting couples: the graph-based approach seems to underestimate the share of couples without children (12.7% EU-SILC vs. 8.4% with graphs) and amplify proportion of families with young children (15.5% vs. 18.9%). Still, we consider the graph-based results as a significant improvement over the original PR families.

24. Starting from the Covid-19 pandemics, the survey mode in EU-SILC and the Labour Force Survey (LFS) has shifted from face-to-face interviews to telephone interviews and Internet. This may weaken the reliability of the place of residence data in the surveys. Another shortcoming of using the survey data is that we observe households as economical units not address-based households. Although these are usually identical, we must acknowledge that survey data does not mimic address-based households perfectly.

Figure 2
Distribution of type of family nucleus from different sources



Sources: EU-SILC 2022, PR – Population Register, Clusters – graph-based approach, after community detection, Final – graph-based approach, after postprocessing.

IV. Conclusion

25. The inaccuracy of place of residence in Estonian PR has an influence on household composition. Families based on PR data are biased towards lone-parent families.

26. The bias of statistics on households and families can be reduced by exploiting other administrative data sources. We considered people and dwellings as nodes of a graph; edges were links found from registers (e.g., marriage connects spouses, property connects apartment with its owner). Then, household and their dwelling could be viewed as a densely

connected subgraph, or in other words, community. To find these, we applied community detection.

27. The resulting statistics align well with family estimates from EU-SILC and are notable advance compared to PR-based statistics.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Gale, D., & Shapley, L. S. (1962). College Admissions and the Stability of Marriage. The American Mathematical Monthly, 69(1), 9. https://doi.org/10.2307/2312726.

Gortfelder, M., & Puur, A. (2021). Tegeliku ja registripõhise elukoha lahknevus ning selle põhjused: 2020. Aasta Eesti tööjõu-uuringu analüüs (lk 31). https://sisu.ut.ee/sites/default/files/mobiilneelu/files/tp1_tlu_tegeliku_ja_registripohise_elukoha_kattuvus_ning_selle_pohjused_etu2020_analuus_gortfelderpuut2021_0.pdf.

Porter, M. A., Onnela, J.-P., & Mucha, P. J. (2009). Communitites in Networks. Notices of the American Mathematical Society, 56(9), 1082–1097.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 105(4), 1118–1123. https://doi.org/10.1073/pnas.0706851105.

Tiit, E.-M., Visk, H., Maasing, E., Levenko, V., & Lehto, K. (2021). Järjekordne rahva ja eluruumide loendus: Milleks ja kuidas? Akadeemia, 2021(11), 2009–2064.

Äär, H. (2017). Coincidence of actual place of residence with Population Register records. Quarterly Bulletin of Statistics Estonia, 1, 80–83.