United Nations ECE/CES/GE.41/2022/6

Economic and Social Council

Distr.: General
23 June 2022

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twenty-fourth Meeting
Geneva, 21−23 September 2022
Item 5 of the provisional agenda
Transitions in census methodology; plans, experiences and innovations

Using Machine Learning Methods to Determine Type of Private Household in the Population and Households Statistics

Note by the Swiss Federal Statistical Office*

Summary

A variable "household type" was created for the Population and Households Statistics (STATPOP). It assigns household type for all private households for the permanent resident population at the main residence. It is mainly based on register information and imputations. This variable has been available since 2010 in the Structural Survey, an annual survey conducted on a sample of about 300,000 individuals which provides direct and reliable estimates for areas of at least 15,000 persons.

This new variable, which is available for all private households of the permanent resident population, makes it possible to conduct analyses at a finer geographic level, such as looking at changes in household types over time in the municipalities. It can also be used at the individual level as an explanatory variable for statistical models and to better understand other topics.

As in the Structural Survey, household type in the Population and Households Statistics (STATPOP) is calculated based on relationships between all household members. However, in the registers, relationships between only some household members are available. The construction of the STATPOP household typology is based on relationship information from registers, supplemented by relationship information from the annual Structural Survey samples. In addition to these two sources, deterministic algorithms are used to identify additional relationships. Together, relationship information from these sources allows us to assign a household type to 87 per cent of households. A Machine

* Prepared by Sabrina Bravo and Estelle Crippa.

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.

learning algorithm is then applied to impute the remaining 13 per cent. Different approaches were tested for imputing missing relationships or missing household types. Based on the performance studies and quality estimates conducted, the decision tree was identified as the best method in this context.

The results are currently being released as an experimental statistic. The household type variable will be integrated into the current STATPOP household production in a second phase. Improvements in imputations and feedback from users are anticipated.

I. Introduction

1. The aim of this project is to create a “household type" variable for private households in Switzerland for the permanent resident population at the main residence. Household type is an important piece of information not only in the household domain but also in relation to other themes. Until 2000, the household type variable was derived from information collected in complete-enumeration population censuses that were held every ten years. With the implementation of the new census system from 2010 onwards, this variable is now available in the Structural Survey, an annual survey carried out on a sample of about 300,000 persons. According to the Structural Survey, household type is based on relationships between all household members. The survey provides direct and reliable estimates for geographic areas of at least 15,000 persons. In the population registers, which are the basis of the population and household statistics, relationships between all household members are not available. Thanks to a new project on household typology, the household type variable is now available in an exhaustive way in the Population and Households Statistics (STATPOP). It allows for analyses at a finer geographic level than the Structural Survey. For example, it is possible to look at the evolution of household types over time in municipalities or to compare work income by household type at a low geographical level. It can also be used at the individual level as an explanatory variable for statistical models to better understand other issues such as poverty analyses.

2. The STATPOP household type variable is created in several steps. First, different data sources are employed to establish the relationships between individuals within each household, then a deterministic algorithm assigns a household type based on the household's relationships for households where all relationships are known. Finally, a machine learning algorithm imputes the household type for households with missing type (approximately 13 per cent).

3. The categories of the household type variable are as follows:[1]

(a) Household of one person;

(b) Married couple without children;

(c) Common-law couple without children[2];

(d) Same-sex couple without children[3];

(e) Lone parent with at least one child under 25 years old[4];

(f) Married couple with at least one child under 25 years old;

(g) Common-law couple with at least one child under 25;

(h) Same-sex couple with at least one child under the age of 25;

(i) Other multi-person household.

[1] A second typology following the same procedure was computed by changing the maximum age considered for children to 18 instead of 25 years old.

[2] For common-law couples, only opposite-sex couples are considered.

[3] Same-sex couples include both common-law couples and registered partnerships.

[4] A child is defined as the daughter or son of a household member.

II. Source of information on relationships

4. Information on the relationships between different household members is taken from the following sources:

(a) Computerized civil status register (INFOSTAR): for all individuals who have had a civil status event in Switzerland (e.g., being born, having a child, getting married, etc.) relationships of "father-child", "mother-child", "wife-spouse" and "registered partner" are recorded. INFOSTAR includes about 85 per cent of the resident population (99.8 per cent for Swiss nationals and 45.5 per cent for foreigners);

(b) Register of diplomats and international organisation staff (ORDIPRO): all individuals with a work permit in Switzerland issued by the Federal Department of Foreign Affairs, as well as their family members and household staff, are recorded in the ORDIPRO register. A variable indicating the relationship between the reference person (diplomat or international organisation staff) and certain adult family members, such as spouse, partner, and child is available in this register;

(c) Structural Survey from 2010 to year n-1: the relationship between each "pair" of household members is derived based on a sample of the population surveyed annually. The relationship reported in the Structural Survey is taken for all pairs of people living together in a household between the time of the survey and the production of the typology, with conditions on marital status for co-habiting individuals, partners, spouses, relatives and non-relatives. All types of relationships that exist in the Structural Survey are included. The Structural Survey publishes its results after the production of the household type variable in STATPOP, which is why the most recent year available is year n-1;

(d) Deterministic algorithms: using demographic variables present in the register-based Population and Households Statistics data (STATPOP), it is possible to define relationships between two persons with a relatively secure result. For instance, two people of different sexes, married with the same date of marriage, living in the same household, will be defined as husband-wife. It is, for example, possible to find married couples who have immigrated and who are not in INFOSTAR data because their civil status event did not take place in Switzerland.

III. Data preparation

5. The relationships taken from the above sources are verified for plausibility (e.g., in a parent-child relationship, the two people must be more than 12 years apart, a child cannot have more than two parents, a person can only be married to one other person) and possible symmetrical relationships are added (in the case where the A-B relationship is described, but not the B-A relationship).

6. Then, distant relationships are added. This consists in deducing, on the basis of the verified relationships, relationships that are not contained in the records (e.g., grandfather, brother-in-law, uncle, etc.). The aim is to identify as many relationships as possible because household type can only be determined if relationships between all household members are known.

7. The contribution of each information source on relationships is, for the year 2018, as follows: of the approximately 16 million "pairs" of people within a household, 84.04 per cent of the relationships are taken from an empirical data source or calculated directly. The relationships are added in the following order: INFOSTAR (75.39 per cent), ORDIPRO (+0.05 per cent), Structural Survey (+6.12 per cent), deterministic algorithm (+1.78 per cent), distant relationships (+0.7 per cent).

Table 1
Available relationships by data source

	INFOSTAR[5]	*ORDIPRO*	*Structural Survey*	*Deterministic algorithm*	*Distant relationships*
Spouse	X	X	X	X	
Registered partner	X	X	X	X	
Co-habiting partner		X	X	X[6]	
Father / mother	X	X	X		
Son / daughter	X	X	X		
Brother or stepbrother / sister or stepsister[7]	X	X	X		
Stepfather / stepmother	X		X		X
Stepson / stepdaughter	X		X		X
Grandfather / grandmother	X	X	X		X
Grandson / granddaughter	X	X	X		X
Related[8]		X	X		X
Unrelated[9]	X		X		X

8. An algorithm that assigns a household type based on the relationships among household members is then applied to households where all relationships are known. In 2018, 13.78 per cent of households have a missing household type, in other words, for 13.78 per cent of households, at least one of its relationships is unknown with the available data and imputation is required.

IV. Imputations

9. Several imputation methods were tested, including random forest on relationships, random forest on household types, a procedure that imputes household type deterministically, multinomial regression, and decision tree. In the end, decision tree was chosen. Although the random forest method gave slightly better individual results, the decision tree method is better at estimating the quality of its imputations and is more flexible for making adjustments.

10. The decision tree is created using a training set that contains known household types and selected variable, which must be available in both the training set and for the set of households to be imputed (e.g., number of persons in the household, average age of the household members, number of different surnames in the household, size of the municipality, etc.). Using statistical tests, the variables that best distinguish the household types are kept and define the selected tree. This tree is then applied to the households requiring imputation, and, depending on the characteristics of these households, a type is imputed.

[5] Only the relationships "Spouse", "Registered partner", "Father/mother" and "Son/daughter" are directly available, the other relationships of the "INFOSTAR" source are inferred.
[6] For couples not married to each other with a common child in the same household.
[7] In the stepbrother/stepsister classification, half-brother and half-sister are also included.
[8] The "related" category includes uncles, aunts, cousins, etc.
[9] The "unrelated" category includes, among others, parents-in-law, children-in-law (children's spouse), children of the co-habiting partner, co-habiting partner of the parent.

11. For the imputations, the Structural Survey plays a key role in creating the training set. Indeed, the intersection between the randomly drawn sample of households in the Structural Survey and the sub-population of households requiring imputation of the same year can be seen as a domain of analysis within the Structural Survey population. This intersection is calculated on the year n-1 and forms the training set for imputing the year n.

V. Results

12. Comparison of distributions for the year 2018: the distribution of household types obtained after imputation on the household population is compared to the distribution of household types estimated by the Structural Survey. The Euclidean distance between these two distributions is 1.92.

Table 2
Distribution of household types for the year 2018 according to the Structural Survey and STATPOP (%)

Household type	*Structural Survey*	*Exhaustive typology (STATPOP)*
Household of one person	35.69	35.69
Lone parent with at least one child under 25	4.64	5.24
Other multi-person household	7.59	8.01
Married couple without children	20.11	19.45
Common-law couple without children	6.53	7.51
Same-sex couple without children	0.61	0.25
Married couple with at least one child under 25 years old	22.12	20.87
Common-law couple with at least one child under 25	2.70	2.96
Same-sex couple with at least one child under the age of 25	0.03	0.03

13. Error estimation for each household type: the intersection between the sample of households from the Structural Survey and the sub-population of imputed households from the same year is used to compare each imputed household type individually. This is possible because the Structural Survey data of that year are not used in the production of the exhaustive "household type"; they can therefore be used as a test set for the imputation model. The Structural Survey weights are taken into account in the error estimates.

Table 3
Estimated imputation error by year

Year	*Percentage of incorrect imputations estimated across all household types*
2018	1.45
2019	1.48
2020	1.47

14. For the moment, two types of errors account for about 76 per cent of the total error (i.e., about 1.1 per cent of the 1.45 per cent). The algorithm does not distinguish properly "Common-law couple without children" with "Other multi-person household" (about 56.5 per cent), and "Same-sex couple without children" with "Other multi-person household" (about 19.5 per cent). This confusion mainly affects two-person households: if the two individuals are of the opposite sex, the algorithm tends to consider them as common-law and if the two persons are of the same sex, the algorithm tends to consider them as roommates. This implies an underestimation of the category "Same-sex couple without children". This is one of the points that will be revised and will significantly improve the quality of imputations.

VI. Conclusion

15. In conclusion, the STATPOP household type variable can be used in a relatively safe way. Indeed, 87 per cent of the household types are obtained on the basis of empirical data and for the remaining 13 per cent of imputed households, the estimated classification error is low (about 1.45 per cent misclassification across all households). Initial feedback from users confirms the benefits of providing this variable in an exhaustive fashion. A few points still need to be improved before this variable can be added to the current production of STATPOP. These include refining the imputation of certain household type categories, adding plausibility checks on household type during imputation,[10] and perhaps grouping certain categories together as long as the variable still meets users' needs. Detailed documentation of the methodology is available on the FSO website (Typology of households (STATPOP)).

[10] Impute the most likely household type according to the decision tree that is also plausible according to the household member data. For example, if "Married couple with at least one child under 25" is the most likely type, but no household member is under 25, the second most likely type, if plausible, will be imputed.