Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO

Household typology in STATPOP statistics (HHTyp)

Production process - short description

Neuchâtel, 2022

Published by:	Federal Statistical Office (FSO)	**Layout concept:**	Section PUB
Information:	info.dem@bfs.admin.ch	**Downloads:**	www.experimental.bfs.admin.ch
Editors:	Sabrina Bravo, FSO; Estelle Crippa, FSO; Fabienne Helfer, FSO	**Copyright:**	FSO, Neuchâtel 2022 Reproduction with mention of source authorised (except for commercial purposes)
Contents:	Sabrina Bravo, FSO; Estelle Crippa, FSO		
Topic:	00 Statistical Basis and Overviews		
Original text:	French		
Translation:	FSO language services		

Introduction

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

The STATPOP variable "household type" has been created in several steps. First, different data sources are employed to establish the relationships between individuals within each household, then an 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 (approx. 13%).

The modalities of the household type variable are as follows:[1]

a) One-person household
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 household with several persons

[1] A second typology following the same procedure was computed by changing the maximum age considered for children of 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.

Source of relationships

Information on the relationships between different individuals in a household is taken from the following sources:

a) *Computerized civil status register* (**INFOSTAR**): in the INFOSTAR data are the "father-child", "mother-child", "wife-spouse" and "registered partner" relationships for all individuals who have had a civil status event (e.g. being born, having a child, getting married, etc.) in Switzerland, which currently represents about 85% of the resident population (99.8% for the Swiss nationals and 45.5% for the 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 giving the relationship between the reference person (diplomat or international organisation staff) and certain adult family members, such as spouse, cohabitee, children, is available in this register.

c) ***Structural Survey*** from 2010 to year n-1: the relationship between each "pair" of household members is surveyed annually directly for a sample of the population. The relationship reported in the Structural Survey is taken for all pairs of people still living in the same household at the time of the production of the typology as at the time of the survey, under the condition that there were no changes in marital status for cohabitants, partners, spouses, relatives and non-relatives. All types of relationships that exist in the Structural Survey are thus included. The Structural Survey publishes its results after the production of the household type variable in STATPOP, which is why the last year available is year n-1.

d) ***Deterministic algorithms***: based on demographic variables present in the register-based Population and Households Statistics (STATPOP) data, 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.

Data preparation

Relationships taken from the above sources are plausibilized (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).

Then, **distant relationships** are added. This consists in deducing, on the basis of the plausibilized relations, relations that are not contained in the records (e.g. grandfather, brother-in-law, uncle, etc.). The aim is to complete the relationships as much as possible, because a household type can only be determined if all the relationships of the household are known.

The contribution of each source of relationships is, for the year 2018, as follows: of the approximately 16 million "pairs" of people within a household, 84.04% of the relationships can be taken from a source at hand or calculated directly. The relationships are added in the following order: INFOSTAR (75.39%), ORDIPRO (+0.05%), Structural Survey (+6.12%), deterministic algorithm (+1.78%), distant relationships (+0.7%).

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	
Cohabitee		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

An algorithm that assigns a household type based on the household member's relationships is then applied and gives a household type to households where all relationships are known. In 2018, 13.78% of households have a missing household type, in other words, for 13.78% of households, at least one of its relationships is unknown with the available data and an imputation is required.

Imputations

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

The decision tree is created using a training set that contains known household types, and selected variables, which must be available in the training set, but also in 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 separate the household types are kept and define the selected tree. This tree is then applied to the households to be imputed and, depending on the characteristics of these households, a type is imputed.

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 to be imputed 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.

---

[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 cohabitee, cohabitee of the parent.

Results

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

Table 2: *Distributions of household types for the year 2018 according to the Structural survey and STATPOP*

Household type	Estimate of the Structural Survey	Exhaustive typology (STATPOP)
One person household	35.69	35.69
Lone parent with at least one child under 25	4.64	5.24
Other household with several persons	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

Individual error estimation: 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 the imputed 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 calculation of the individual error estimate on the household table.

Table 3: *Estimated individual imputation error by year*

Year	Percentage of incorrect imputations estimated on total households
2018	1.45%
2019	1.48%
2020	1.47%

For the moment, two types of errors account for about 76% of the total errors (i.e., about 1.1% of the 1.45%). Indeed, the algorithm does not distinguish properly "Common-law couple without children" with "Other household with several persons" (about 56.5%), and "Same-sex couple without children" with "Other household with several persons" (about 19.5%). 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.

Conclusion

In conclusion, the STATPOP household type variable can be used in a relatively safe way; indeed, 87% of the household types are obtained on the basis of collected data and for the remaining 13% of imputed households, the estimate of the classification error is low (about 1.45% of misclassification on all households). Initial users' feedbacks confirm the benefits of having 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 categories. Detailed documentation of the methodology is available on the FSO website (Typology of households (STATPOP)).