Machine Learning Methods for the Imputation of the Type of Private Household

UNECE Expert Meeting on Population and Housing Censuses

September 21-23, 2022

Dr. Jean-Paul Kauthen, Swiss Federal Statistical Office

Work by

- Sabrina Bravo
- Estelle Crippa

Outline of the presentation

1. Context
2. Motivation
3. Production and imputation
4. Results
5. Conclusion

1. Context

- Until 2000, the information on the type of private household was collected every 10 years in the full enumeration population census.
- Since 2010, the information on type of private household is collected in the Structural Survey, an annual sample of 300,000 persons (yields reliable estimations for groups > 15,000 persons).
- Since 2021, this type is produced with both the information available in the registers and **machine learning** algorithms. Publication on the FSO's experimental statistics website.

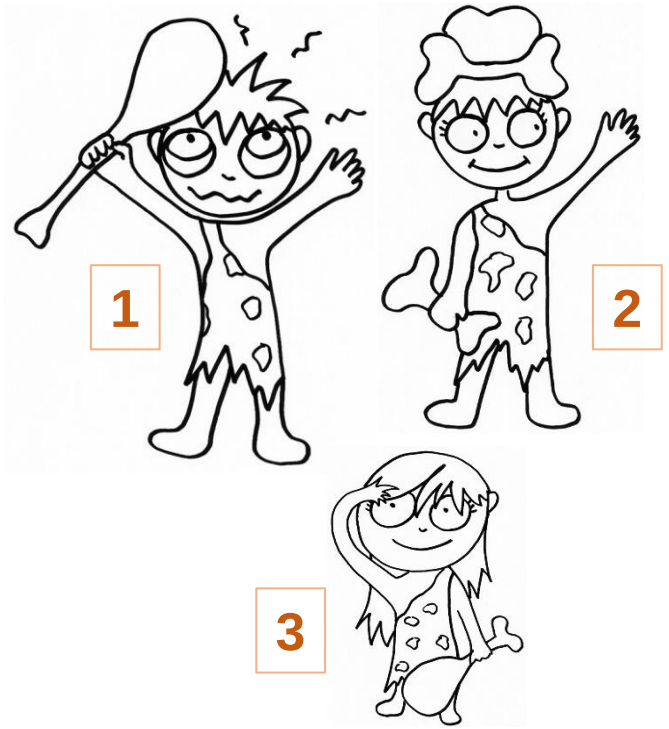
2. Motivation

Exhaustive information on type of household allow for:

- Analysis at a fine geographical level, for example:
  - Distribution of household types in different neighborhoods of large cities;
  - Municipality-level comparison of the social assistance rate, by household type;
- Studies at the individual level as an explanatory variable for statistical models, such as poverty analysis.

3. Production - process

The household type is based on the relationships in the household



Relationships

P1	P2	Relationship
1	2	Husband
1	3	Father
2	1	Wife
2	3	Mother
3	1	Daughter
3	2	Daughter

Household type

Married couple with child

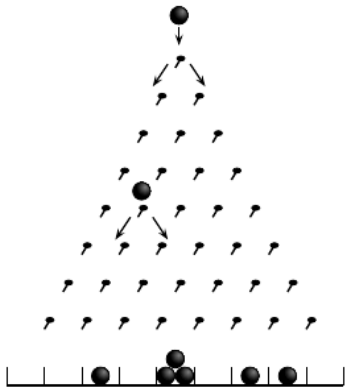
3. Production - sources

Data sources for the household type :

	Relationships	Household type
1. Relationships taken from the *registers*	75.4%	
2. Relationships taken from the *Structural Survey*	+6.1%	
3. Relationships deduced by *deterministic algorithms*	+2.5%	
	84% →	86.2%
4. Types imputed with ***machine learning*** (decision tree)		+13.8%
		100%

3. Production - imputation

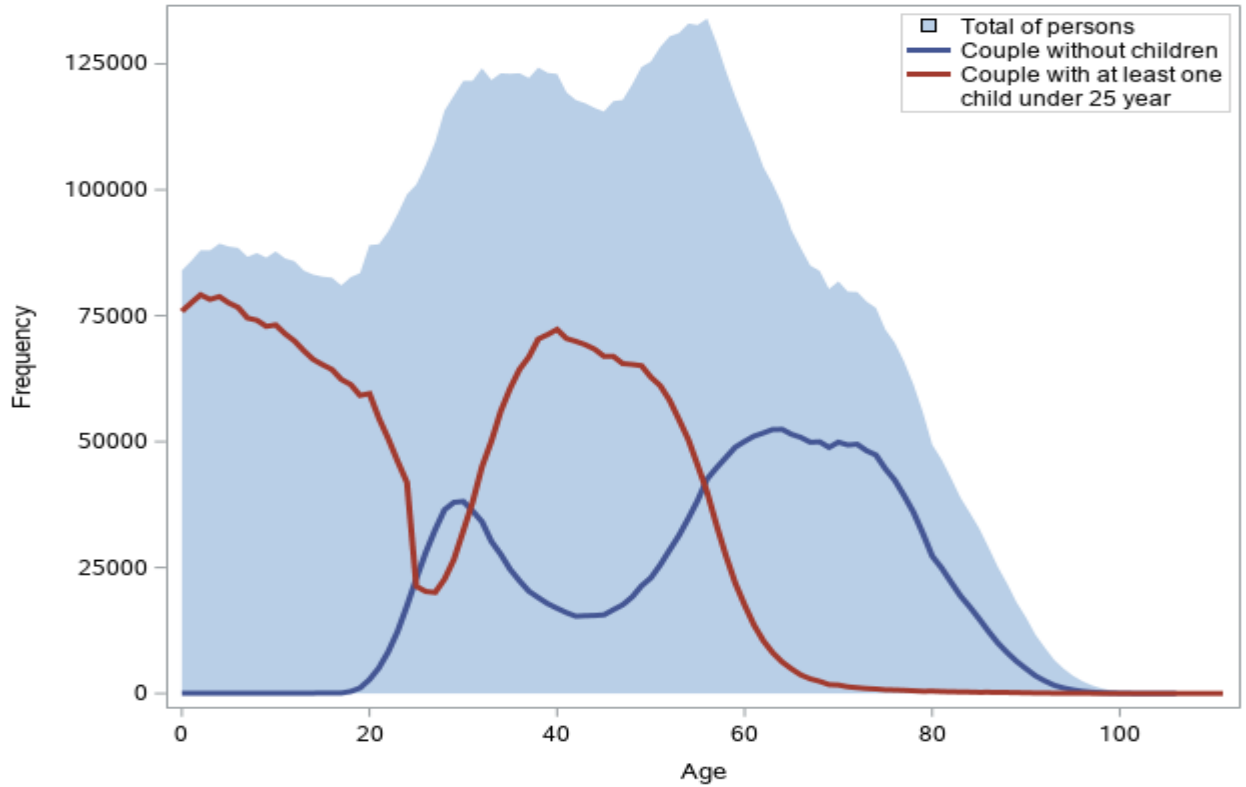
- Different imputation methods were tested for the 14% of households whose type could not be defined because at least one household relationship is unknown.
- **The decision tree** was finally chosen as the method for imputing the remaining households:



1. Use of a training set that contains households of known type and selected variables (e.g. average age of household members, size of municipality, ...).
2. By means of statistical tests on the training set, the variables that best separate the household types are retained and define the selected tree.
3. This tree is then applied to the households to be imputed and, depending on the characteristics of these households, a type is imputed.

4. Results - example of dissemination

Permanent resident population in private households by age and household type, 2020



5. Conclusion

- 86% of household types are obtained on the basis of available data and 14% with machine learning.
- Approximately 1.45% of misallocations across all households.
- Initial feedback from users confirms the need for this variable to be exhaustive, i.e. no missing values and available for the whole resident population.
- A few points will be improved before adding this variable in the current annual production.