Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

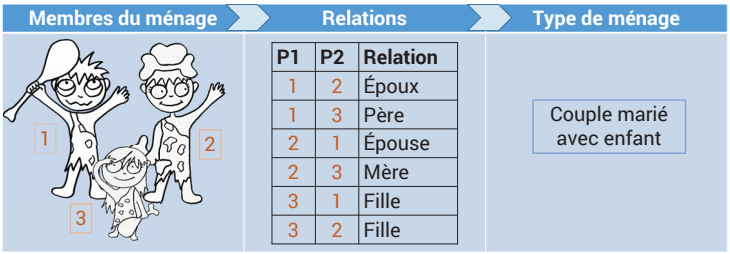
Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Uffizi federal da statistica UST

Utilisation de méthodes de Machine Learning pour déterminer le type des ménages privés dans la Statistique de la Population et des Ménages

Introduction Depuis 2022, l'OFS diffuse sur son site des statistiques expérimentales une variable « Type de ménage » pour l'ensemble des ménages privés de la population résidante permanente au domicile principal. Depuis 2010, cette variable est disponible dans le Relevé Structurel, une enquête par échantillonnage. Cette nouvelle variable exhaustive de la Statistique de la Population et des Ménages (STATPOP) permet de faire des analyses pour des groupes plus petits que le Relevé Structurel (min. 15'000 personnes), comme par exemple, de regarder les évolutions des types de ménage à travers le temps dans les communes. Elle peut également être utilisée au niveau individuel comme variable explicative pour des modèles statistiques et ainsi aider à mieux comprendre des problématiques concernant différentes thématiques.

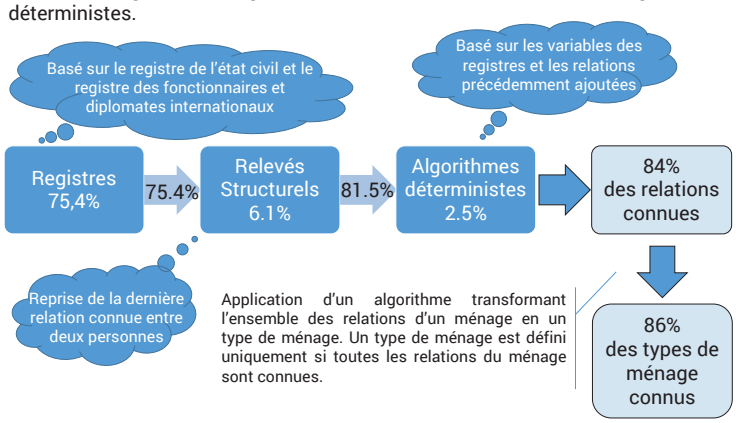
Procédure

Pour chaque ménage, les relations entre les personnes permettent de définir un type de ménage.



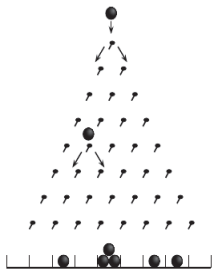
Sources des relations

Trois types de source nous informent sur les relations entre les personnes d'un même ménage : les registres, le Relevé Structurel et des algorithmes déterministes.



Machine Learning

Pour imputer les 14% des types de ménage manquants, plusieurs méthodes d'imputation ont été testées, notamment un random forest sur les relations, un random forest sur les types de ménage, une procédure qui impute le type de ménage de manière déterministe, une régression multinomiale et un arbre de décision. Bien que le random forest donnait des résultats individuels légèrement meilleurs, l'arbre de décision a été retenu, car il estime mieux la qualité de ses imputations et est plus flexible pour faire des ajustements.



1. Utilisation d'un ensemble d'apprentissage qui contient des ménages dont le type est connu et des variables choisies (p. ex. moyenne d'âge des membres du ménage, taille de la commune, …).
2. Au moyen de tests statistiques sur l'ensemble d'apprentissage, les variables qui séparent au mieux les types de ménage sont gardées et définissent l'arbre retenu.
3. Cet arbre est, dans un deuxième temps, appliqué aux ménages à imputer et selon les caractéristiques de ces ménages, un type est imputé.

Performance des imputations

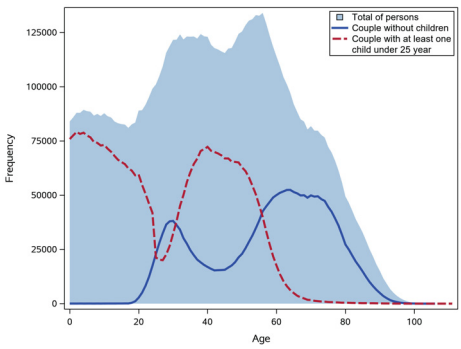
Comparaison des distributions pour l'année 2018 : la distribution obtenue après imputation sur la population des ménages est comparée à la distribution estimée par le Relevé Structurel. La distance euclidienne entre ces deux distributions est de 1,92.

Type de ménage	Estimation du Relevé Structurel	Typologie exhaustive (STATPOP)
Ménage d'une personne	35,69	35,69
Parent seul avec au moins un enfant de moins de 25 ans	4,64	5,24
Autre ménage de plusieurs personnes	7,59	8,01
Couple marié sans enfant	20,11	19,45
Couple en union libre sans enfant	6,53	7,51
Couple de même sexe sans enfant	0,61	0,25
Couple marié avec au moins un enfant de moins de 25 ans	22,12	20,87
Couple en union libre avec au moins un enfant de moins de 25 ans	2,70	2,96
Couple de même sexe avec au moins un enfant de moins de 25 ans	0,03	0,03

Estimation des erreurs individuelles : les données du Relevé Structurel de l'année en cours ne sont pas utilisées dans la production du type de ménage exhaustif. Elles peuvent donc être utilisées pour estimer le pourcentage d'imputations incorrectes.

Année	Erreurs estimées sur le total des ménages
2018	1,45 %
2019	1,48 %
2020	1,47 %

Exemples d'utilisation de la variable « Type de ménage »



Comparaison du nombre de personnes avec les types de ménage « couple avec enfant de moins de 25 ans » et « couples sans enfant » selon l'âge, en 2020.

Taux de ménages avec le type « parent seul avec au moins un enfant de moins de 25 ans » par commune, en 2020.

Conclusion L'ensemble des ménages privés se voit attribuer un type de ménage, 86% proviennent de données relevées et 14% sont imputés par machine learning, avec une estimation de mauvaises attributions d'environ 1,45% sur l'ensemble des ménages. Les premiers retours d'utilisateurs confirment la nécessité d'avoir cette variable de manière exhaustive. Le « Type de ménage » sera ajouté à la production courante STATPOP après quelques améliorations.

Projekt/Thema: Typologie des ménages (STATPOP)
Anlass: Statistiktage 2022
Datum: 26. – 27.10.2022

Autor/in: Sabrina Bravo
Organisationseinheit: Sektion POP
sabrina.bravo@bfs.admin.ch

© BFS/OFS, Neuchâtel 2022
https://www.experimental.bfs.admin.ch/
www.statistik.ch