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

Conseil économique et social

Distr. générale
23 juin 2022
Français
Original : anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Groupe d'experts des recensements de la population et des habitations

Vingt-quatrième réunion
Genève, 21-23 septembre 2022
Point 5 de l'ordre du jour provisoire
Transitions dans les méthodes appliquées pour les recensements : plans, expérience acquise et innovations

Utilisation de méthodes d'apprentissage automatique pour déterminer le type de ménage privé dans les statistiques de la population et des ménages

Note de l'Office fédéral de la statistique suisse*

Résumé

Une variable « type de ménage » a été créée pour les statistiques sur la population et les ménages (STATPOP). Elle sert à attribuer le type de ménage pour tous les ménages privés de la population résidente permanente à la résidence principale. Elle est principalement basée sur les informations des registres, complétées par des imputations. Cette variable est disponible depuis 2010 dans le relevé structurel, une enquête annuelle menée sur un échantillon d'environ 300 000 personnes qui fournit des estimations directes et fiables pour les zones d'au moins 15 000 personnes.

Cette nouvelle variable, qui est disponible pour tous les ménages privés de la population résidente permanente, permet de réaliser des analyses à un niveau géographique plus fin, comme l'examen de l'évolution des types de ménages au fil du temps dans les communes. Elle peut également être utilisée au niveau individuel comme variable explicative pour des modèles statistiques et pour mieux comprendre d'autres sujets.

* Document établi par Sabrina Bravo et Estelle Crippa.
Note : Les appellations employées dans le présent document ne reflètent aucune prise de position du Secrétariat de l'Organisation des Nations Unies quant au statut juridique de pays, territoires, villes ou zones quelconques, ou de leurs autorités, ni quant au tracé de leurs frontières ou limites.

Comme dans le relevé structurel, le type de ménage dans les statistiques sur la population et les ménages (STATPOP) est calculé sur la base des relations entre tous les membres du ménage. Cependant, dans les registres, seules sont disponibles les relations entre certains membres du ménage. La construction de la typologie des ménages dans les STATPOP est basée sur les informations sur les relations provenant des registres, complétées par celles provenant des échantillons du relevé structurel annuel. En plus de ces deux sources, des algorithmes déterministes sont utilisés pour dégager des relations supplémentaires. Ensemble, les informations sur les relations provenant de ces sources nous permettent d'attribuer un type de ménage à 87 % des ménages. Un algorithme d'apprentissage automatique est ensuite appliqué pour imputer les 13 % restants. Différentes approches ont été testées pour l'imputation des relations manquantes ou des types de ménages manquants. Sur la base des études de performance et des estimations de qualité réalisées, il a été déterminé qu'un arbre de décision était la meilleure méthode dans ce contexte.

Les résultats sont actuellement publiés en tant que statistiques expérimentales. C'est dans une deuxième phase que la variable du type de ménage sera intégrée dans la production actuelle des STATPOP concernant les ménages. Il est prévu d'améliorer les imputations et de tenir compte des retours d'utilisateurs.

I. Introduction

1. Le but du projet décrit dans le présent document est de créer une variable « type de ménage » pour les ménages privés en Suisse pour la population résidente permanente à la résidence principale. Le type de ménage est un élément d'information important, non seulement pour l'étude des ménages eux-mêmes mais aussi en relation avec d'autres thèmes. Jusqu'en 2000, la variable relative au type de ménage était dérivée des informations recueillies lors des recensements de la population par dénombrement complet qui avaient lieu tous les dix ans. Avec la mise en place du nouveau système de recensement à partir de 2010, cette variable est désormais disponible dans le relevé structurel, une enquête annuelle réalisée sur un échantillon d'environ 300 000 personnes. D'après le relevé structurel, le type de ménage est basé sur les relations entre tous les membres du ménage. Cette enquête fournit des estimations directes et fiables pour les zones géographiques d'au moins 15 000 personnes. Cependant, dans les registres de la population, qui constituent la base des statistiques de la population et des ménages, toutes les relations entre membres du ménage ne sont pas disponibles. Grâce à un nouveau projet sur la typologie des ménages, la variable « type de ménage » est maintenant disponible de manière exhaustive dans les statistiques sur la population et les ménages (les STATPOP). Elle rend possible des analyses à un niveau géographique plus fin que le relevé structurel. Par exemple, il est possible d'examiner l'évolution dans le temps des types de ménages au sein des municipalités ou de comparer les revenus du travail par type de ménage à un niveau géographique détaillé. Elle peut également être utilisée au niveau individuel comme variable explicative pour des modèles statistiques afin de mieux comprendre d'autres problématiques, par exemple pour analyser la pauvreté.

2. La variable « type de ménage » dans les STATPOP est créée en plusieurs étapes. Tout d'abord, différentes sources de données sont utilisées pour établir les relations entre les individus au sein de chaque ménage, puis un algorithme déterministe attribue, pour les ménages dont toutes les relations sont connues, un type de ménage sur la base des relations du ménage. Enfin, un algorithme d'apprentissage automatique impute le type de ménage pour les ménages dont le type est manquant (environ 13 %).

3. Les catégories de la variable « type de ménage » sont les suivantes[1] :

a) Ménage d'une personne ;

b) Couple marié sans enfant ;

c) Couple en union libre sans enfant[2] ;

d) Couple de même sexe sans enfant[3] ;

e) Parent seul avec au moins un enfant de moins de 25 ans[4] ;

f) Couple marié avec au moins un enfant de moins de 25 ans ;

g) Couple en union libre ayant au moins un enfant de moins de 25 ans ;

h) Couple de même sexe avec au moins un enfant de moins de 25 ans ;

i) Autre ménage composé de plusieurs personnes.

[1] Une deuxième typologie suivant la même procédure a été calculée en changeant l'âge maximum considéré pour les enfants à 18 au lieu de 25 ans.

[2] Pour les couples en union libre, seuls les couples de sexe opposé sont pris en compte.

[3] Les couples de même sexe comprennent à la fois des couples en union libre et des couples en partenariat enregistré.

[4] Un enfant est défini comme la fille ou le fils d'un membre du ménage.

II. Sources d'information sur les relations

4. Les informations sur les relations entre les différents membres du ménage proviennent des sources suivantes :

a) Registre informatisé de l'état civil (INFOSTAR) : pour tous les individus qui ont eu un événement d'état civil en Suisse (par exemple, naissance, enfant, mariage, etc.), les relations « père-enfant », « mère-enfant », « épouse-conjoint » et « partenaire enregistré » sont enregistrées. INFOSTAR comprend environ 85 % de la population résidente (99,8 % pour les citoyens suisses et 45,5 % pour les étrangers) ;

b) Registre des diplomates et du personnel des organisations internationales (ORDIPRO) : toutes les personnes titulaires d'un permis de travail en Suisse délivré par le Département fédéral des affaires étrangères, ainsi que les membres de leur famille et le personnel de maison, sont inscrits dans le registre ORDIPRO. Une variable indiquant la relation entre la personne de référence (diplomate ou fonctionnaire d'une organisation internationale) et certains membres adultes de la famille, tels que conjoint, partenaire et enfant, est disponible dans ce registre ;

c) Relevé structurel de 2010 à l'année n-1 : la relation entre chaque « paire » de membres du ménage est dérivée sur la base d'un échantillon de la population visée par l'enquête chaque année. La relation indiquée dans le relevé structurel est prise pour toutes les paires de personnes vivant ensemble dans un ménage entre le moment de l'enquête et la production de la typologie, avec des conditions sur la situation matrimoniale pour les individus cohabitants, les partenaires, les conjoints, les membres de la parentèle et les non-parents. Tous les types de relations qui existent dans le relevé structurel sont inclus. Les résultats du relevé structurel sont publiés après la production de la variable « type de ménage » dans STATPOP, c'est pourquoi l'année disponible la plus récente est l'année n-1 ;

d) Algorithmes déterministes : en utilisant les variables démographiques présentes dans les données statistiques sur la population et les ménages basées sur les registres (les STATPOP), il est possible de définir les relations entre deux personnes et de parvenir à un résultat relativement sûr. Par exemple, deux personnes de sexe différent, mariées avec la même date de mariage, vivant dans le même ménage, seront définies comme mari et femme. Il est, par exemple, possible de trouver des couples mariés qui ont immigré et qui ne figurent pas dans les données d'INFOSTAR parce que leur événement d'état civil n'a pas eu lieu en Suisse.

III. Préparation des données

5. La plausibilité des relations tirées des sources citées ci-dessus est vérifiée (par exemple, dans une relation parent-enfant, les deux personnes doivent avoir plus de 12 ans de différence, un enfant ne peut pas avoir plus de deux parents, une personne ne peut être mariée qu'à une seule autre personne) et les éventuelles relations symétriques sont ajoutées (dans le cas où la relation A-B est décrite, mais pas la relation B-A).

6. Ensuite, les relations plus lointaines sont ajoutées. Cela consiste à déduire, sur la base des liens de parenté vérifiés, des liens de parenté qui ne sont pas indiqués dans les registres (par exemple, grand-père, beau-frère, oncle, etc.). L'objectif est de définir autant de relations que possible, car le type de ménage ne peut être déterminé que si les relations entre tous les membres du ménage sont connues.

7. La contribution de chaque source d'information sur les relations est, pour l'année 2018, la suivante : sur les quelque 16 millions de « paires » de personnes au sein d'un ménage, 84,04 % des relations sont issues d'une source de données empiriques ou calculées directement. Les relations sont ajoutées dans l'ordre suivant : INFOSTAR (75,39 %), ORDIPRO (+0,05 %), relevé structurel (+6,12 %), algorithme déterministe (+1,78 %), relations lointaines (+0,7 %).

Tableau 1
Relations disponibles par source de données

	INFOSTAR[5]	*ORDIPRO*	*Relevé structurel*	*Algorithme déterministe*	*Relations lointaines*
Conjoint(e)	X	X	X	X	
Partenaire enregistré(e)	X	X	X	X	
Partenaire cohabitant(e)		X	X	X[6]	
Père/mère	X	X	X		
Fils/fille	X	X	X		
Frère ou quasi-frère/sœur ou quasi-sœur[7]	X	X	X		
Beau-père/belle-mère	X		X		X
Beau-fils/belle-fille	X		X		X
Grand-père/grand-mère	X	X	X		X
Petit-fils/petite-fille	X	X	X		X
Membre de la parentèle[8]		X	X		X
Sans lien de parenté[9]	X		X		X

8. Un algorithme qui attribue un type de ménage en fonction des relations entre les membres du ménage est ensuite appliqué aux ménages où toutes les relations sont connues. En 2018, 13,78 % des ménages ont un type de ménage manquant, c'est-à-dire que pour 13,78 % des ménages, au moins une de ses relations est inconnue avec les données disponibles et une imputation est nécessaire.

IV. Imputations

9. Plusieurs méthodes d'imputation ont été mises à l'essai, notamment la forêt aléatoire sur les relations, la forêt aléatoire sur les types de ménage, une procédure qui impute le type de ménage de manière déterministe, la régression multinomiale et l'arbre de décision. C'est en définitive l'arbre de décision qui a été choisi. Bien que la méthode de la forêt aléatoire ait donné des résultats individuels légèrement meilleurs, la méthode de l'arbre de décision permet de mieux estimer la qualité des imputations et offre davantage de souplesse pour faire des ajustements.

10. L'arbre de décision est créé à l'aide d'un ensemble de données d'apprentissage contenant des types de ménages connus et une variable sélectionnée, qui doit être disponible à la fois dans l'ensemble d'apprentissage et pour l'ensemble des ménages nécessitant une imputation (par exemple, le nombre de personnes dans le ménage, l'âge moyen des membres du ménage, le nombre de noms de famille distincts dans le ménage, la taille de la municipalité, etc.). À l'issue de tests statistiques, les variables qui distinguent le mieux les types de ménages sont conservées et définissent l'arbre sélectionné. Cet arbre est ensuite appliqué aux ménages nécessitant une imputation et, suivant les caractéristiques de ces ménages, un type est imputé.

11. Pour les imputations, le relevé structurel est une base essentielle pour la création de l'ensemble de données d'apprentissage. En effet, l'intersection entre l'échantillon de ménages tiré au hasard dans le relevé structurel et la sous-population de ménages nécessitant une imputation pour la même année peut être considérée comme un domaine d'analyse au

[5] Seules les relations « Conjoint(e) », « Partenaire enregistré(e) », « Père/mère » et « Fils/fille » sont directement disponibles, les autres relations de la source « INFOSTAR » étant déduites.
[6] Pour les couples non mariés l'un à l'autre ayant un enfant commun dans le même ménage.
[7] Dans la catégorie des quasi-frères et quasi-sœurs, on inclut également les demi-frères et demi-sœurs.
[8] La catégorie « membre de la parentèle » comprend les oncles, les tantes, les cousins, etc.
[9] La catégorie « sans lien de parenté » comprend, entre autres, les beaux-parents, les beaux-enfants (conjoints des enfants), les enfants du partenaire cohabitant, le partenaire cohabitant du parent.

sein de la population du relevé structurel. Cette intersection est calculée sur l'année n-1 et constitue l'ensemble de données d'apprentissage pour les imputations de l'année n.

V. Résultats

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

Tableau 2
Distribution des types de ménages pour l'année 2018 selon le relevé structurel et les STATPOP (en %)

Type de ménage	*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 composé 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 ayant 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

13. Estimation des erreurs pour chaque type de ménage : l'intersection entre l'échantillon de ménages du relevé structurel et la sous-population de ménages ayant fait l'objet d'une imputation la même année est utilisée pour comparer un par un chaque type de ménage imputé. Cela est possible parce que les données du relevé structurel de cette année-là ne sont pas utilisées dans la production de la typologie exhaustive ; elles peuvent donc être utilisées comme ensemble de test pour le modèle d'imputation. Les pondérations du relevé structurel sont prises en compte dans les estimations d'erreurs.

Tableau 3
Estimation des erreurs d'imputation, par année

Année	*Pourcentage d'imputations incorrectes estimé pour tous les types de ménages*
2018	1,45
2019	1,48
2020	1,47

14. Pour l'instant, deux types d'erreurs représentent environ 76 % du total des erreurs (c'est-à-dire environ 1,1 % de l'ensemble des imputations, par rapport à 1,45 % pour le total des erreurs). L'algorithme ne distingue pas correctement les « couples en union libre sans enfant » des « autres ménages composés de plusieurs personnes » (environ 56,5 %), ni les « couples de même sexe sans enfant » des « autres ménages composés de plusieurs personnes » (environ 19,5 %). Cette confusion touche principalement les ménages de deux personnes : si les deux individus sont de sexe opposé, l'algorithme a tendance à les considérer comme un couple en union libre, et si les deux personnes sont de même sexe, l'algorithme a tendance à les considérer comme des colocataires. Cela entraîne une sous-estimation de la catégorie « Couple de même sexe sans enfant ». C'est un des points qui feront l'objet d'une révision, et cela devrait améliorer considérablement la qualité des imputations.

VI. Conclusion

15. En conclusion, on peut dire que la variable « type de ménage » des STATPOP peut être utilisée de manière relativement sûre. En effet, 87 % des types de ménages sont obtenus sur la base de données empiriques et, pour les 13 % restants de ménages pour lesquels le type est imputé, l'erreur de classement estimée est faible (environ 1,45 % pour l'ensemble des ménages). Les premiers retours des utilisateurs confirment l'intérêt de fournir cette variable de manière exhaustive. Quelques points doivent encore être améliorés avant que cette variable puisse être ajoutée à la production actuelle des STATPOP. Il s'agit notamment d'affiner l'imputation de certaines catégories de type de ménage, d'ajouter des contrôles de plausibilité sur le type de ménage au moment de l'imputation[10], et peut-être de regrouper certaines catégories tant que la variable répond toujours aux besoins des utilisateurs. Une documentation détaillée sur la méthodologie est disponible sur le site de l'OFS (Typologie des ménages (STATPOP)).

[10] Imputer le type de ménage le plus probable selon l'arbre de décision qui est également plausible d'après les données des membres du ménage. Par exemple, si « Couple marié avec au moins un enfant de moins de 25 ans » est le type le plus probable, mais qu'aucun membre du ménage n'a moins de 25 ans, le deuxième type le plus probable, s'il est plausible, sera imputé.